1989

Long-term consistency and stability of some neuropsychological measures with normal and disabled readers.

Michael C. S. Harnadek
University of Windsor

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

LONG-TERM CONSISTENCY AND STABILITY OF SOME NEUROPSYCHOLOGICAL MEASURES WITH NORMAL AND DISABLED READERS

by

Michael C. S. Harnadek

B. Sc. University of Victoria, 1987

A Thesis
Submitted to the Faculty of Graduate Studies
through the Department of Psychology
in Partial Fulfillment of the
Requirements for the Degree
of Master of Arts at the
University of Windsor
Windsor, Ontario, Canada
1989

ISBN 0-315-54501-1

Canada

ACM6128

ABSTRACT

The present study is divided into four investigations. Each examined the test-retest reliability of neuropsychological measures used in the assessment of children. Reliability is assessed using two indices, consistency (Pearson-r correlation) and internal stability (intraclass correlation). Tests were divided into the following ability domains that they are thought to subserve: Psychometric Intelligence, Motor, Academic Achievement and Reading, Auditory-Perception and Language, Tactile-Perception, Visual-Perception, Right-Left Awareness, and the Underlining Test (Doehring, 1968). Investigation 1 is a retrospective examination of the two-year retest reliability of 25 disabled readers' (DR) performance on a series of neuropsychological measures. These children displayed poor to fair consistency ($r < .40$) and poor stability (ICC < .29) across the majority of the test measures. The most consistent retest performance is observed for eight measures of Motor ability. Two measures of each, Motor and Visual-Perception ability are associated with the largest degree of interal stability. Investigation 2 examined the retest performance of a matched sample (grade, sex, IQ, retest interval) of 27 normal readers (NR) on the same test

measures. The largest degree of consistency and stability is associated with measures of Academic Achievement and Reading, and of Psychometric Intelligence. In comparison to the DR group, a greater degree of reliability is seen for the NR childrens' retest performance. The two groups' rank orders of stability and consistency coefficients were not found to be significantly similar. Effects due to sampling bias, maturation change and development, treatment and experiential factors, and possible ceiling effects in some of the data are reasons forwarded to account for the DR sample's lower reliability. Investigation 3 compares the reliability of the DR group's performance with that of two heterogeneous clinical samples of children. While the DR group demonstrates a lesser degree of reliability than the clinical samples, their patterns of consistency and stability were significantly similar. The results of Investigations 1 through 3 suggest that while the long-term retest reliability of the DR sample is more variable than either normal reading or heterogeneous clinical samples of children the pattern of reliability across measures resembles that of the clinical samples. Investigation 4 attemptes to validate the retest reliability of 42 child subjects pooled together. The pooled sample was tested twice over a two-year interval (Year0-Year2), and again a further two-years later (Year2-Year4). The rank orders of stability and consistency generated for the initial two-year

retest period are compared with those for the subsequent retest interval. The second retest pattern of consistency was found to be signicantly similar to the initial pattern, however the meaningfulness of the comparison is thrown into doubt due to the large number of non-significant correlation coefficients.

ACKNOWLEDGEMENTS

Completion of this thesis, and the initial two years of graduate school during which time it was written, has been exhausting, sometimes troubling, and always challenging. But the challenges have resulted in a great many personally rewarding experiences. Throughout this period several people have contributed to the rewards that I have experienced, and are deserving of my gratitude.

First and foremost, I want to thank my parents for their continuous encouragement and never ending belief in me. Graduate students do not just "happen". The necesary attitudes, beliefs, goals, and dedication that have culminated in the successful completion of this thesis have been nurtured in me from a very early age. As such, this work is as much an extension of my parents as it is of myself. I dedicate this thesis to them.

The members of this thesis committee, Drs. Shore and Williams deserve thanks for the time, effort, and understanding that they expended in helping me successfully complete this project.

I would like to extend my special gratitude to Dr. Byron Rourke, the chairperson of this thesis, and my advisor in graduate school. Both student and advisor share a common

goal, the making of a clinical neuropsychologist, and both must trust the other's judgement and motivation. I have been entirely satisfied with the guidance and training that I have received from Dr. Rourke over the past two years and I am grateful for the opportunity to learn under his tutelage.

Finally, I am thankful for the tremendous understanding, support, and counselling that Gloria has provided to me. During this past year she has had to endure many sacrifices so as to allow me to complete this work. I regret that such sacrifices had to be made, but there is no one else that I would have wanted by my side throughout this experience.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Test-retest Reliability

The reliability of a test, or a device, is an estimate of its consistency in measurement. Nunnally (1978) refers to reliability as a measure of the repeatability of a score. Expressed in terms of a correlation coefficient (Anastasi, 1982), the degree of agreement between sets of scores, resulting from two applications of the device under similar testing conditions, is revealed by the magnitude of the coefficient. Anastasi (1982) describes various forms of reliability that are available to describe a test's consistency: Test-retest reliability (temporal stability), alternate-forms reliability (consistency between forms), split-half reliability and inter-item reliability (internal consistency), and inter-rater reliability (consistency between judges).

Test-retest reliability refers to the degree of generalization that can be made of a test's scores over specified lengths of time. Stated more specifically, test-retest reliability is concerned with the extent to which an individual's performance on a test, over more than one occasion, reflects the "true" ability under investigation,

rather than error variance. Probable sources of errorvariance include carry-over effects from repeated administrations of the test, and situational differences between test administrations not directly related to the test content (Anastasi, 1982). Inter-judge variability can also constitute an additional source of error affecting test-retest correlations (Nunnally, 1978).

Sources of error can affect correlation coefficients in a variety of different manners. Practice and memory effects can spuriously increase the value of a resultant correlation coefficient. Conversely, reliability coefficients can be reduced through the negative influence of fatigue upon the examinee (Lord & Novick, 1968).

The length of time between administrations of a test can contribute to either increasing or decreasing the resulting correlation coefficients between the sets of scores. With longer time intervals, changes in test performance are likely to be progressive and cumulative, reflecting intra-individual changes as well as changes in the ability under consideration (Anastasi, 1982; Lord & Novick, 1968). The possibility of real change occurring in the observed behaviour increases with longer retest intervals (Sechrest, 1984). Longer periods of time separating the testing sessions can also result in reduced agreement between scores because of decreased memory effects (Lord & Novick, 1968). Shorter periods of intervening time

can provide spuriously high test-retest reliability coefficients through positive memory and practice effects (Lord & Novick, 1968), or through the subject's grasping of the nature of the test (Anastasi, 1982).

Test-retest reliability involves calculation of a correlation coefficient between the scores of the same test, administered under the same testing conditions, and separated by varying lengths of time (Anastasi, 1982). The critical element in the reliability calculation may be reduced to the time interval separating testing sessions by controlling the testing contents (through the use of identical forms), utilizing the same individuals, and by equating the testing conditions as much as possible.

In classical test theory, the comparison of scores is commonly expressed in the form of a Pearson product-moment correlation coefficient (Pearson-r) the computation of which is based upon the standard error of the measurement (Anastasi, 1982; Lindquist, 1953). Alternatively, test-retest reliability may be conceptualized within the model of generalization theory (Cronbach, Rajaratnum, & Gleser, 1963). Generalization theory differs from classical test theory by incorporating the use of the Intraclass correlation coefficient (ICC) as the measure of test-retest reliability. Subsequent discussion will focus upon Pearson-r and ICC as these pertain to test-retest reliability. The advantages and disadvantages associated with each measure

will be commented upon, and the basic composition of the Pearson-r and the ICC will be discussed.

Classical Test Theory and Pearson Product-Moment Correlation

Within the realm of classical test theory, a subject's score on a test is assumed to be composed of a true score, characteristic of the behaviour of interest, plus some random error (Sechrest, 1984). Coefficients of reliability are commonly expressed in terms of Pearson-r (Anastasi, 1982). Designed for use with bivariate data distributions, the Pearson-r incorporates the position of the individual's test score within the distribution, and its deviation above and below the mean, into the calculation of the correlation (Anastasi, 1982). Assumptions requiring fulfilment for use of this statistic include equality of variance and intercorrelations (Cronbach, Ikeda, & Avner, 1962; Cronbach, et al., 1963). The mean of the products between the standardized scores of two measures provides the correlation coefficient, and indicates the degree of agreement between the measures (Anastasi, 1982; Haggard, 1958). This manner of computation for Pearson-r has been cited as a shortcoming in its use in reliability studies (Bartko, 1976; Haggard, 1958).

Employing Pearson-r to calculate test-retest reliabilities for tests involves certain limitations. As was previously mentioned, the manner in which the Pearson-r is calculated restricts the type of information that is made

available to the researcher. Standardization of the scores during calculation of the Pearson-r necessitates equalizing the means and variances of the scores (Haggard, 1958). This procedural step precludes the researcher from examining any intra-individual differences that may be of interest to him or her (Haggard, 1958). Intra-individual variance is pooled with error variance in calculation of the correlation coefficient, and only the variance between the factors of interest are examined for their degree of relatedness (Sechrest, 1984; Winer, 1971). When the reliability of a test is thought to be dependent upon more than one source of variance, the Pearson-r method of correlation will prove inadequate in providing information concerning the influence of the additional sources of variation (Berk, 1979; Haggard, 1958). As stated by Lindquist (1953), the use of the classical method of reliability coefficients is "decidedly inadequate ... when several sources of random error may be distinguished" (p. 357).

Related to the above-mentioned limitations, and perhaps the single most detrimental limitation of the Pearson-r method, is its poor degree of internal stability. As described by several authors (Bartko, 1976; Bartko & Carpenter, 1976; Haggard, 1958) this method of correlation demonstrates a positive additive and multiplicative bias. According to Barko and Carpenter (1976) the covarying of scores in such a manner so as to retain their overall

pattern, will fail to produce changes in correlations based upon the Pearson-r method. Employing an example based upon Table 3 of Bartko and Carpenter (1976), one set of scores (1, 2, 3, 4, 5 and 1, 2, 3, 4, 5) will yield a correlation coefficient that is identical to a coefficient calculated upon a biased second set of scores (1, 2, 3, 4, 5 and 2, 4, 6, 8, 10). In both of the above cases, the Pearson-r coefficient is 1.0; however, the systematic additive bias of the second set is hidden from the investigator if sought by use of the Pearson-r method. The above limitations of the Pearson-r method of test-retest reliability have resulted in the advocacy of an alternative measure of reliability based upon generalizability theory (Bartko, 1976; Berk, 1979; Haggard, 1958; Lindquist, 1953). The next section of this paper will discuss generalizability theory and the Intraclass correlation coefficient.

Generalizability Theory and Intraclass Correlation

Generalizability theory (Cronbach, et al., 1963; Sechrest, 1984) conceptualizes a subject's score on a test as being a sample drawn from a pre-described universe of possible scores on that particular measure. The universe of scores from which the sample is drawn is described in terms of the facets that cause its make-up (Sechrest, 1984). Generalizability theory offers a useful alternative to the classical method in calculating test-retest reliability coefficients. While the Pearson-r coefficient illuminates a

single source of variance - differences between the standardized means of each testing situation (Anastasi, 1982; Haggard, 1958), the intraclass correlation coefficient is sensitive to multiple sources of variance that give rise to the observed score (Sechrest, 1984). Flexibility in examining one or more individual sources of variances permits the researcher greater precision in generalizing the reliability to a specified population (Cronbach, et al., 1963).

In calculating reliability coefficients within generalizability theory, the ICC is used. Based upon analysis of variance components, the ICC can be calculated for individual sources of variance present in the reliability model (Haggard, 1958). The value of the ICC is expressed as the ratio of the amount of variance of interest to the amount of variance of interest combined with residual error variance (Bartko, 1966; Lahey, Downey & Saal, 1983). Specifying individual sources of variance not of interest can allow those sources to be partialled out of the total error variance, resulting in a more precise estimate of the reliability (Haggard, 1958). The ICC is described as being a measure of the homogeneity of scores within a testing situation in relation to the total variance present (Haggard, 1958).

Consistency between scores and internal stability make the ICC a useful complement to the Pearson-r. Consistency

of differences between score means over two similar testing situations is provided by both measures. Internal stability of test scores, the similarity of the score means over assessment situations, is only addressed by the ICC (Bartko & Carpenter, 1976; Brown, 1985). The ICC is sensitive to differences in the means and variances of the scores being collected, with larger differences resulting in a reduced magnitude for the ICC (Haggard, 1958). The advantages of employing ICC in studies of reliability are discussed by Berk (1979). The ICC provides an estimate of reliability that is free of additive or multiplicative bias. A comprehensive analysis of the factors related to the estimate of reliability is possible as are decisions concerning the inclusion or exclusion of different sources of variance when calculating the reliability coefficient. A more precise estimate of the reliability coefficient is thus attainable. The ICC is a suitable statistic to employ with a variety of categorical and quantitative sources of observation. Finally, the ICC is able to treat designs of unequal sample size with ease.

Test-retest Reliability: Standards of Acceptability

In discussing test-retest reliability, some manner of interpreting the value of the resultant correlation coefficients, in terms of their strength of agreement, is necessary. In addition, interpretations must be made in reference to the length of intervening time span involved

between administrations of the test. As indicated by Anastasi (1982), the length of time dividing test administrations shares a negative relationship with the value of the resultant test-retest reliability correlation coefficient.

As a measure of the stability of individual traits, correlation coefficients in excess of .50 are regarded by Bloom (1964) as being within the realm of acceptability. In making his recommendation, however, Bloom was referring to an assessment interval of one year or greater (Bloom, 1964). In contrast to the above suggestion for stability, the consistency of test measurements over time are recommended to have much greater values if they are to be considered acceptable. Berk (1978) recommends that high degrees of consistency are indicated by correlation coefficients in the .80s. This interpretation of retest correlation coefficients is shared by Anastasi (1982). A more conservative value of correlation coefficient is suggested by Bloom in regard to assessment of a test's consistency. Bloom (1964) recommends correlation coefficient values equal to or in excess of .85 if interpretations of strong agreement are to be made.

Cicchetti and Sparrow (1981) suggested a set of criteria for evaluating the strength of relationship conveyed by a correlation coefficient. In their study, coefficients that were less than .39 in value were rated to

have poor predictive usefulness. Coefficients ranging from .40 to .59 were judged to have fair retest reliability. Coefficients ranging in value from .60 to .74, and .75 and greater, were deemed to be good and excellent, respectively predictors of future behavior. However, as Brown, Rourke, and Cicchetti (1989) point out, the criteria forwarded by Cicchetti and Sparrow (1981) is more commonly applied to shorter retest intervals (less than six months). For longer retest periods, Brown et al. (1989) suggest that more liberal criteria should be adopted. Specificly, correlations suggestive af excellent or good reliability would be associated with coefficients greater than .60, and from .40 to .59, respectively. Fair reliability would be indicated by a coefficient ranging in value from .30 to .39. Finally, coefficient values less than .30 would be afforded a poor rating. For the proposes of this study the criteria presented by Brown et al. (1989) will serve as the measures against which the clinical significance of a long-term retest coefficient is judged.

Examination of past studies that addressed the retest reliability of some or all of the measures employed in the present investigation can provide a useful framework within which the emergent results may be interpreted. While a sufficient number of studies exist that examine this issue with measures of psychometric intelligence and of academic achievement, few studies have directly addressed the issue

of test-retest reliability with neuropsychological measures. A summary of the past research that has attempted to identify the test-retest reliability of various neuropsychological measures will be presented first.

The dearth of studies of this type within the literature makes it necessary to discuss both those that have employed adults as subjects and those that have utilized child subjects. However, these two groupings of studies will be discussed separately. By including past studies involving adult subjects, it is hoped that some interpretive feeling is provided to the reader with which he can supplement the studies that employed children.

Discussion of the retest reliability studies involving tests of psychometric intelligence, the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) and the Peabody Picture Vocabulary Test (PPVT; Dunn, 1965) follows that of the neuropsychological measures. Finally, past studies employing the measures of academic achievement and reading ability, the Wide Range Achievement Test (WRAT; Jastak & Jastak, 1965) and the Metropolitan Achievement Test (MAT; Durost, Bixler, Wrightstone, Prescott, & Balow, 1971), will be discussed. Throughout this discussion, the terms 'test-retest reliability' and 'retest reliability' will be used synonymously.

Test-retest Reliability of the Halstead-Reitan Neuropsychological Test Battery

Of the tests comprising the Halstead-Reitan Neuropsychological Test Battery (Reitan, 1969; Reitan & Wolfson, 1985), the most commonly employed measures studied in the literature are the following: Category Test (Cat); Tactual Performance Test (TPT-Time, TPT-Location, TPT-Memory); Finger Tapping Test (FTT); Seashore Rhythm Test (SRT); Speech-Sounds Perception Test (SSPT); Grip Strength (Grip) and the Trail Making Test for Adults (TMT, Forms A and B).

The first systematic examination of long-term test-retest reliability of some of the subtests comprising the Halstead-Reitan Neuropsychological Test Battery (HRNTB) was conducted by Klonoff, Fibiger, and Hutton (1970). In their study, Klonoff, et al. examined the 12-month retest reliability of six subtests (Cat, TPT, SRT, SSPT, FTT, TMT) administered to 35 chronic schizophrenic patients. Considering the psychometric consistency of the subtest measures, Klonoff, et al. report that only two measures (TPT-Location, TPT-Memory) achieved correlation coefficient values less than .70 (Klonoff, et al., 1970). The authors summarized their findings as having demonstrated adequate one-year test-retest reliability of these measures for this clinical sample (Klonoff, et al., 1971). The Pearson-*r* correlation coefficients resulting from this study are

included in Appendix A.

In a later study, Matarazzo, Wiens, Matarazzo, and Goldstein (1974) investigated the clinical and psychometric retest reliability of some neuropsychological measures in a sample of normal adult males. The sample employed in the study consisted of 29 neurologically sound adult males, with a mean age of 24-years. The test-retest interval utilized in the study was 20-weeks. The resultant Pearson-r correlation coefficients for the measures examined (Cat, TPT, SRT, SSPT, FTT, TMT, Grip) ranged from .24 (FTT) to .68 (TPT-Time). The results of Matarazzo, et al. (1974) are included in Appendix B for comparison purposes with those of Klonoff, et al. (1971), and with other studies still to be discussed.

The results of the normal adult male sample were compared with those from a second sample consisting of 16 elderly patients (mean age = 60-years) previously diagnosed as having diffuse cerebrovascular disease. Mean test-retest interval for this second sample was 12.4-weeks. Retest reliabilities for the same neuropsychological measures yielded a range of values from .44 (FTT) to .96 (Cat). Five of the 11 measures examined attained correlation coefficients greater than .70 (Cat, TPT-Time, TPT-Memory, TPT-Location, TMT A).

Interpretation of the results of the Matarazzo, et al. (1974) study indicates that a greater degree of variability

may be seen in the performance scores of adult normals over a five month period than can be expected for an elderly clinical population.

In an extension of the previously cited study by Matarazzo, et al. (1974), Matarazzo and colleagues (Matarazzo, Matarazzo, Wiens, Gallo, & Klonoff, 1976) examined the 20 week retest reliability of the HRNTB on a sample of 15 carotid endarterectomy patients. Due to the fact that all of these patients underwent surgery for their medical condition between neuropsychological assessments, it is probable that this confound precludes accurate assessment of the measures' test-retest reliability. Drawing comparisons between the results for this sample and those from three previously reported samples (Klonoff, et al., 1970; Matarazzo, et al., 1974), the authors report adequate reliability for most of the test measures employed (Matarazzo, et al., 1976). Pearson-r correlation coefficients ranged from .30 (Non-dominant Grip) to .93 (TPT-Time); and, eight of the 10 measures yielded correlations in excess of .70 (Cat, TPT-Time, TPT-Location, SRT, SSPT, TMT A, TMT B, & FTT). These results are presented in Appendix C.

The FTT was examined by Morrison, Gregory, and Paul (1979) to determine its various reliability coefficients. Included within this study was an examination of the test-retest reliability of the FTT using a sample of 60 young

adults (Modal age = 19-years). Although it is not clear if the authors were concerned with the initial neurological status of the sample, one can make the assumption that none suffered severly disabling neurological impairment since all subjects were recruited from college-level courses. Two assessments were performed, with an average retest interval of one week intervening (Morrison, et al., 1979). Pearson-r correlations were calculated for both the dominant hand (.80) and the non-dominant hand (.82). Employing a standard of .80 for correlation coefficients to be considered as demonstrating adequate reliability, the authors conclude that the FTT is barely adequate as a reliable measure (Morrison, et al., 1979).

Dodrill and Troupin (1975) compared the retest reliabilities of four administrations of some of the HRNTB subtests with a sample of 17 chronic epileptic adults (mean age = 27.41-years). Retest intervals ranged from six to 12 months, with the total study period not exceeding 29 months for any of the subjects. All possible administrations were compared to one another to determine if the resultant correlation coefficients would be reduced following repeated administrations. Pearson-r correlation coefficients were calculated for each comparison. Possible contamination of the second administration occurred due to 9 of the 12 subjects having received treatment with the drug Sulthiame during this time. [In reference to the adverse effects of

Sulthiame on neuropsychological test performance, Green, Troupin, Halpern, Friel, and Kanarek (1974) reported that subjects who were administered this medication were less alert, and had demonstrated reduced performance on timed tasks.] For the results of the Dodrill and Troupin (1975) study, practice effects were only seen for two measures (Cat, TPT-Location). Highest reliability was evident in two measures of motor performance (FTT, Grip). Three measures demonstrated a relative decrease in reliability with increasing retest intervals (TPT-Memory, TMT A, TMT B), while four measures remained relatively the same in terms of their reliability coefficients (TPT-Time, SSPT, FTT, Dominant Grip). Three other measures (Cat, TPT-Location, SRT) demonstrated variability in coefficient values, with no firm pattern established. In general, the HRNTB subtests demonstrated adequate reliability on this small sample of epileptic patients.

Eckardt and Matarazzo (1981) initiated a study of a heterogeneous sample of males with mild to moderate neuropsychological impairment. The test-retest reliability of nine measures (Cat, TPT-Time, TPT-Location, TPT-Memory, SRT, SSPT, FTT, TMT A, TMT B) was compared between a sample of 91 drug-free alcoholic inpatients, and 20 non-alcoholic medical inpatients (Eckardt & Matarazzo, 1981). The mean retest intervals differed between the two samples: that of the alcoholic sample 16.8 days, while the medical sample

averaged 22.9 days between test administrations. The reliability coefficients (Pearson-r) were generally higher for the medical patients; coefficients ranged from .51 (TPT-Memory) to .94 (TMT B). Six of the nine measures (Cat, TPT-Time, SSPT, FTT, TMT A, TMT B) exceeded a coefficient value of .80. None of the reliability coefficients for the alcoholic patients exceeded .80, and values ranged from .53 (TPT-Location, SRT, TMT A) to .74 for the Category Test (Eckardt & Matarazzo, 1981). All results are included in Appendix D.

Eckardt and Matarazzo interpreted their findings in context with those of Dodrill and Troupin (1975); specifically that patients with mild to moderate neuropsychological impairments may demonstrate test-retest reliabilities unique to this population (Eckardt & Matarazzo, 1981).

Test-retest Reliability of Neuropsychological Tests Used in the Assessment of Children

As previously mentioned, few studies of the test-retest reliability of neuropsychological assessment measures with children have been conducted. Furthermore, of the few studies that have been completed (Brown, 1987; Donaghy, 1988; Knights & Moule, 1967; Paniak, 1987; Sarazin & Spreen, 1986; Russell & Rourke, 1984), only the study by Sarazin and Spreen investigated the test-retest reliability of a variety of neuropsychological measures in a relatively distinct

clinical sample.

Knights and Moule (1967), in addition to providing normative data on finger tapping and foot tapping performance in children, examined the test-retest reliability of these two measures. The samples employed in the study were comprised of two groups of elementary school children. The first sample consisted of 72 children with no history of academic failure. Thirty-one children referred for neurological evaluation because of academic and/or behaviourial problems formed the second sample. The range of ages for all children was from approximately 5- to 15-years. The range of intelligence quotients for these subjects was from 67 to 126 (mean IQ = 102.3).

The children in the first sample were retested after 12 months, and the scores of the two trials were compared using a Pearson-r correlation coefficient. Correlation coefficient values for the non-dominant hand (r = .88), and non-dominant foot (r = .85) were higher than those for the dominant hand (r = .78) and dominant foot (r = .81). Interpretation of the results of this sample would indicate acceptable levels of test-retest reliability for both measures with normal children.

The second sample of children was re-assessed after an average of seven months on the two tapping tasks (Knights & Moule, 1967). Although the specific correlation coefficients were not provided by the authors, Knights and

Moule (1967) describe the results are having a pattern very similar to that of the first sample.

Sarazin and Spreen (1986) conducted a 15-year longitudinal study of the stability of some neuropsychological measures with a sample of children referred for assessment because of learning disability. The 175 children who comprised the sample were divided into three clinical diagnostic groups: Brain Damaged (BD), Minimal Brain Dysfunction (MBD), and Learning Disabled (LD; Sarazin & Spreen, 1986).

The long-term consistency of test measures was calculated through use of Pearson-r correlation coefficients. Coefficients were highest for measures of lateral dominance; these were followed by tests of psychometric intelligence, academic achievement, and language. Low values of test-retest reliability were reported for measures of concept formation and motor ability. A significant difference in the performance of all three groups over time suggested low internal stability for the neuropsychological measures, despite adequate psychometric retest reliability, over the 15-year study period (Sarazin & Spreen, 1986). Additionally, the length of the retest interval necessitated the use of adult forms for two of the measures investigated (WAIS-R and modified adult version of the Category Test). This would seem to have restricted the generalizability of results to this

particular sample.

A longitudinal study investigating the predictive and concurrent validity of phonetic accuracy in learning disabled children was carried out by Russell and Rourke (1984). Twenty-one children, ranging in age from 9 to 14 years, were assessed over a 12-month period. All were referred for assessment because of suspected learning disabilities, and each was of normal intelligence. A wide range of variables were employed in this study. Measures of psychometric intelligence, personality, and behaviour were assessed twice, interspaced by a 12-month retest period. Measures of achievement, language, visual-spatial skills, and simple motor skills were assessed four times during the 12-month extent of the study (i.e., once every three months). Test-retest reliabilities were determined for the childrens' phonetic accuracy and for each of the neuropsychological measures correlated with phonetic accuracy.

A detailed examination of the reliability and validity of the Halstead-Reitan neuropsychological tests with children was conducted by Brown (1987). Two hundred and forty-eight children were assessed twice. The measures involved tests of learning, behaviour, or perception in which possible cerebral involvement was suspected (Brown, 1987). The children had a mean age of 8-years (SD = 1.7) at the time of the first assessment. Composition of the

clinical sample revealed that the subjects fell within one or more of five clinical classifications: learning disabled, mentally retarded, brain-damaged, emotionally disturbed, and/or environmentally deprived. The reader is referred to Brown (1987) for a more detailed discussion of the criteria used to classify children into the clinical categories.

The tests employed by Brown (1987) assessed a full range of behaviourial and ability domains. Brown (1987) categorized these abilities under the following headings: Psychometric Intelligence, Academic Achievement, Motor ability, Auditory-Verbal skill, Visual-Spatial ability, Sequencing skill, and Tactile-Perception. Some of the test measures employed in Brown (1987) are directly relevent to the present study; these are presented in a different organizational schema in Appendix E. The retest interval separating the two assessments was on the average, 2.65-years (SD = 1.74). In view of the range of ages for the sample and the lengthy retest interval, only those tests that were administered in the identical manner to all subjects were included in the analyses (Brown, 1987).

A series of 50 Pearson-r correlation coefficients were calculated, providing an assessment of each measure's test-retest consistency. The method of ICC was employed to provide an evaluation of each measure's internal stability over time. Appendix E contains consistency and stability

coefficients for all of the measures investigated in Brown (1987).

Considering the consistency of measures over time, the range of the resultant correlation coefficients was from .02 (Auditory (L) errors) to .82 (Full-Scale IQ). The highest reliability coefficients were yielded by the summary indices of the WISC: Full-Scale IQ, Verbal IQ, and Performance IQ had Pearson-r correlations of .75 or greater.

The coefficients of reliability for measures of Academic Achievement were all moderate. Wide Range Achievement Test subtest reliability coefficients ranged from .51 (Arithmetic) to .58 (Spelling). A more detailed presentation of the WRAT correlations is presented in the section of this introduction devoted to the discussion of that test measure.

Within the domain of Language abilities, most of the measures studied demonstrated moderate consistency over time. The most reliable measures proved to be the tests of Sentence Memory and PPVT-IQ. Both tests yielded coefficients of reliability of .71. Measures of Speech - Sounds Perception Test (correct) yielded the lowest reliability of the Language tests ($r = .47$).

Brown (1987) separated subjects into right- and left-hand dominant subgroups when discussing those tests that she had classified as falling within Motor or Sensory ability domains. Generally, moderate to low consistency was evident

in the performance of right-hand dominant children on the Sensory measures. Coefficients of reliability ranged from .14 for Visual (L) errors to .48 for TPT Location correct. When examining the consistency of this same sub-sample on the Motor measures, slightly greater reliability coefficients are evident. The most reliable measure was Maze Time (L), yielding a correlation coefficient of .67. The lowest degree of retest reliability within the Motor domain was for the measure of Name Writing speed (R)[.18].

Those children identified as being left-hand dominant yielded moderate to low consistency in performances on the Sensory tests. The coefficients ranged in value from .69 (Finger Agnosia (L) errors), to .02 (Auditory (L) errors). Those tests falling within the Motor domain demonstrated slightly greater test-retest reliability than did the Sensory measures for right- and left-hand dominant children.

Overall, the majority of neuropsychological and other test measures exceeded a Pearson-r correlation value of .50. The highest consistency was measures of Psychometric Intelligence, followed by those measures in the Language domain, and Motor tests. The lowest degrees in consistency were evident for the Sensory tests.

Referring to the stability of these same measures over time, we must examine the ICC coefficient values of Brown (1987). For the most part, the ICC values very closely resembled the pattern of the Pearson-r correlation

coefficients. There was a general tendency for the ICCs to be lower than the Pearson-r coefficients by a margin of .01 to .02. However, there were comparisons in which substantial discrepancies occurred between consistency and stability coefficient values. Such was the case for the Motor tests of Grip (R) and Grip (L) assessed within both handedness subgroups, and in the measures of Tactile (R) errors and Tactile (L) errors, Visual (R) errors, and Auditory (R) errors for left-hand dominant subjects.

As in the case of consistency, the measures with the greatest stability appear to be the summary indices of Psychometric Intelligence (Full-Scale IQ, Verbal IQ, Performance IQ), and two measures of Language skill (Sentence Memory, PPVT-IQ). The lowest degree of stability was generally associated with the Motor and Sensory measures.

If one were to employ the standard suggested by Brown et al. (1989), in which coefficient values greater than or equal to .30 would indicate fair or better long-term stability, the majority of the Brown (1987) tests would succeed.

Paniak (1987), in a direct replication of the Brown (1987) study, examined the test-retest reliability of a similar set of neuropsychological test measures. He employed a unique sample of 75 children drawn from the same population as was used in Brown (1987). [The reader is

referred to the discussion of Brown (1987) for a description of the diagnostic categories used in the Paniak (1987) study.] The initial age of the sample had a mean of 7.8-years (SD = 1.9). Mean age of the sample after the retest assessment was 10.3-years (SD = 2.3).

The sample was divided into the same diagnostic categories employed in the Brown (1987) study. Criteria for the inclusion of subjects into the various diagnostic categories can be found in Paniak (1987) or Brown (1987).

The average retest interval for Paniak (1987) was 2.49-years (SD = 1.73). The tests of interest examined a broad range of abilities. A list of the tests used in the Paniak (1987) study can be found in Appendix E as can the test-retest correlation coefficients that resulted.

While none of the 50 Pearson-r reliability coefficients calculated as greater than .80 for any of the measures assessed, excellent reliability (r > .60) was seen for four measures of Psychometric Intelligence (Performance IQ, Full-Scale IQ, Object Assembly, Block Design), for two measures of Motor ability (Foot Tapping - right and left foot), and for two measures of Language ability (Auditory Closure, PPVT-IQ). An additional acceptable degree of reliability was demonstrated for the Target Test. Reliability coefficients less than .30 resulted for one of the 14 measures of Psychometric Intelligence, for 1 of 14 measures of Motor ability, and for seven of 13 measures of sensory

perception. The complete results of the reliability comparisons performed in Paniak (1987) are provided in Appendix E.

As mentioned within the discussion of the Brown (1987) study, the heterogeneous nature of the sample employed in the Paniak (1987) study precludes generalization of the findings to finely specified populations. However, some comparison can be made between the Paniak (1987) results and those of some of the studies previously discussed. The test-retest reliability coefficients that emerged for the finger- and foot-tapping tests of Knights and Moule (1967) are considerably greater than those in the Paniak (1987) study. This discrepancy can largely be attributed to differences in the retest intervals (12-months in Knights & Moule, 1987 versus an average of 2.49-years in the Paniak, 1987 study).

Comparison with the Brown (1987) study illustrates some similar findings between those of the Paniak (1987) study. Suitable degrees of test consistency and stability were demonstrated for the Full-Scale and Performance IQ from the WISC, the PPVT-IQ, and the Target Test. Low correlations were yielded by tests of psychomotor ability (Name Writing-Time, left and right hand), sensorimotor tests (TPT-Time both hands, TPT-Location, TPT-Memory), and a measure of psychometric intelligence (Comprehension subtest of WISC) in both studies.

In order to expand the findings of the Paniak (1987) and Brown (1987) studies and to apply them to larger heterogenous clinical populations, Donaghy (1988) combined the two previous samples into a single investigation. The number of subjects comprising this combined sample was 322 children. the mean age of the sample was eight-years (SD = 1.80) at the time of their first assessment, and 10.6-years (SD = 2.30) at the completion of the study. The average test-retest interval in Donaghy's investigation was 2.62-years (SD = 1.73). The characteristics of the sample of children employed included a majority of males (79.5 percent) and of right-hand dominant cases (82.5 percent at the time of the second assessment).

Identical diagnostic classification categories to those used in the studies by Brown (1987) and Paniak (1987), were employed in Donaghy (1988). (The reader may wish to refer to the earlier description of these studies in order to review how the children were diagnostically classified.)

Summarizing the findings of Donaghy (1988), the Pearson-r correlation coefficients closely resembled the pattern identified in the Brown (1987) results. The highest degree of consistency over time was evident in the summary indices of the WISC. Full-Scale IQ provided a coefficient of .81, and the reliability coefficients for Verbal IQ and Performance IQ were equal to or greater than, .76. The smallest degree of reliability evident within the

Psychometric Intelligence domain was for the Comprehension subtest; it was associated with a retest reliability coefficient of .45. High to moderate coefficients were reported for the Language measures. Values ranged from .70 (PPVT-IQ) to .48 for Speech-Sounds Perception Test correct. The three WRAT scores comprising the Achievement domain exhibited moderate retest reliability coefficients. Values ranged from .56 (WRAT Reading) to .50 (WRAT Arithmetic).

Separate analyses were performed for right- and left-hand dominant children on the Motor and Sensory measures. For the right-hand dominant subgroup, the reliability of the Motor tests was moderate to low. A coefficient of .66 was reported for Maze Time (R), Maze Counter (R) errors, and Maze Time (L). A Pearson-r coefficient of .17 was yielded by the Name (R) speed measure. Left-hand dominant children provided somewhat greater coefficients for the Motor tests. A Pearson-r value of .74 was reported for Foot Tapping (R). the lowest coefficient of reliability was .18 for Name (R) speed.

The Sensory tests generally proved to be less consistent over time than the Motor tests. A range of values from .69 (Finger Agnosia (L) errors) to .08 (Auditory (L) errors) was provided for the left-hand dominant subsample. The Sensory measures of right-handed children also proved to be moderate to low in terms of their consistency over time. A moderate value of .48 was reported

for TPT Location correct; and, a low value of .12 was yielded by Auditory (L) errors.

The pattern of retest reliability coefficients between the six domains used in Donaghy (1988) was similar to those revealled in the results of Brown (1987). Donaghy reports that such a finding is not particularly surprising given the large proportion of subjects (77 percent) in this study that had originally been employed in the Brown (1987) study.

Test-retest Reliability of the WISC Variables

The Wechsler Intelligence Scale for Children (WISC) has enjoyed a comparatively greater degree of research that has examined its associated psychometric properties than have the neuropsychological tests previously discussed. Extensive reviews of the reliability and validity of the WISC with a variety of populations are provided in Littell (1960), and more recently in Zimmerman and Woo-sam (1972).

Long-term Test-retest Reliability

Eight studies were reviewed that examined the long-term stability of the WISC in association with so-called normal children (Conklin & Dockrell, 1967; Gehman & Matyas, 1956), mentally retarded children (Friedman, 1970; Reger, 1962; Whately & Plant, 1957), learning-disabled children (Coleman, 1963), and heterogeneous child clinical samples (Brown, 1987; Donaghy, 1988; Paniak, 1987). In general, the Full-Scale IQ, Verbal IQ, and Performance IQ summary indices of the WISC proved to be reliable to an acceptable degree

(Pearson-r coefficient > .70) in the majority of these studies.

An early study into the long-term stability of the WISC was initiated by Gehman and Matyas (1956). The test-retest reliability of the three summary indices of the WISC (Full-Scale IQ, Verbal IQ, Performance IQ) were evaluated over a four-year retest interval. Sixty children of normal intellectual level, having a sample mean age of 15-years and 2-months, served as the study's subjects (Gehman & Matyas, 1956). Correlation coefficients of .77 resulted between pre- and post-comparisons of both Verbal IQ and Full-Scale IQ. The correlation coefficient for Performance IQ was .74. There was some increase in these three summary scores over the four-year period, but the increases were non-significant, suggesting an adequate degree of stability for these measures (Gehman & Matyas, 1956).

Conklin and Dockrell (1967) studied the longitudinal prediction of academic achievement from children's WISC scores. In the process, these authors were able to provide two- and four-year test-retest reliability data on the WISC. From an original sample of 101 children (mean age = 10-years, 5-months), 70 children remained who had received two assessments with the WISC after a two-year retest interval. A third assessment followed a further two-year retest period, during which time the study's sample had been reduced to 60 children for the four-year comparison (Conklin

& Dockrell, 1967).

The consistency of the summary indices was evaluated for both, two- and four-year test comparisons. The two-year comparison yielded retest reliability coefficients of above .70 for the Full-Scale and Verbal IQ summary indices, and a coefficient in the high .50s for the Performance IQ summary score (Conklin & Dockrell, 1967). The four-year stability coefficients ranged from .54 (Performance IQ) to .72 (Full-Scale IQ), resembling the results of the two-year evaluation (Conklin & Dockrell, 1967). Conklin and Dockrell interpreted the test-retest reliability of the Verbal and Full-Scale IQ indices as being in keeping with the results of Gehman and Matyas (1956).

The 10 subtests evaluated over the four-year interval (Digit Span and Mazes subtests were not included) generally proved to be less reliable over multiple administrations than the summary indices. This pattern of results is not uncommon in the literature, and has been reported in the past (Conklin & Dockrell, 1967). The test-retest reliability coefficients showed a large degree of variability, ranging from .13 (Picture Arrangement) to .59 for the Coding subtest (Conklin & Dockrell, 1967).

Studies by Friedman (1970), Reger (1962), and Whately and Plant (1957) provide an assessment of the long-term stability of the WISC as it is used with mentally retarded children. Whately and Plant (1957) examined the stability

of the WISC with 70 elementary-aged children, all of whom had initial Full-Scale IQ standard scores of 90 or lower. The retest interval was at least 12-months for all children, with the median number of months intervening equalling 17. Comparison of the three IQ summary scores over the two administrations revealled no significant differences (Whately & Plant, 1957). Although no coefficients of reliability were provided, the authors interpret their results as having demonstrated the stability of the Full-Scale, Verbal, and Performance IQ summary indices with this clinical sample (Whately & Plant, 1957).

Reger (1962) included 65 educably mentally retarded boys as the sample for his study. Each of the youths was administered the WISC three times, with a mean retest interval of 12-months separating each assessment. As in the Whately and Plant (1957) study, no coefficients of reliability were provided. Instead, two-tailed *t*-tests were employed, the results of which indicated significant improvements ($p < .01$) had occurred in the mean Full-Scale IQ and Performance IQ standard scores over time, demonstrating reduced consistency for these measures (Reger, 1962). The Verbal IQ summary score remained reliable through out all three administrations with this sample (Reger, 1962).

Coleman (1963) studied the performance of a group of 24 learning disabled boys (mean age = 11.4-years) in his

investigation of the long-term test-retest reliability of this measure with a sample of learning disabled children. Three of the children had a WISC Full-Scale IQ of below 90; however, the group was considered to be of normal intellectual level for the purposes of the study (Coleman, 1963). The period of time between WISC administrations was on average 15-months.

The reliability of the summary indices of the WISC that emerged from this study were compared to that which was evident in the Gehman and Matyas (1956) study. The Full-Scale IQ reliability was equal to that seen in the earlier (Gehman & Matyas, 1956) study; however, the reliability of the Verbal IQ score was considerably lower (.62 versus .77). Performance IQ reliability (.81) was higher than that seen in the comparison study. The author accounts for the difference in the reliability of the two Verbal IQ summary scores by suggesting that the learning disabled sample may have had lower motivation on those subtests that resembled school-type (mainly verbal) learning (Coleman, 1963).

Test-retest reliability for the WISC subtests ranged from poor (Comprehension, Pearson-r = .29) to excellent (Object Assembly, Pearson-r = .68), yielding only three correlation coefficients greater than .60 (Vocabulary, Coding, Object Assembly). When interpreting his findings in light of those that resulted from the study by Gehman and Matyas (1956), Coleman (1963) suggested that a greater

degree of variance in the retest reliability may be present in a learning disabled population than in a non-disabled population of children.

The recent studies by Brown (1987), Paniak (1987), and Donaghy (1988) provided an indication of the stability of heterogeneous clinical samples' performance on the WISC. The reader is referred to the earlier discussion of these three studies for information concerning the composition of the samples used, and the procedures employed in them. The results of these investigations as they pertain to the test-retest reliability of the WISC is provided below.

Using an average retest interval of 2.65-years, Brown (1987) found quite acceptable degrees of test-retest reliability for the WISC summary scores for her clinical sample. Pearson-r coefficients were reported to be .82 for the Full-Scale and Verbal IQ indices, and .75 for the Performance IQ summary score (Brown, 1987). There was some slight variability in the reliability of the subtests over this time period; however, all proved to have a good level of consistency. The range of the Pearson-r coefficients extended from .46 (Comprehension) to .66 for Vocabulary (Brown, 1987).

Paniak (1987) found a somewhat different pattern in the reliabilities for the summary indices of Intelligence. Evaluated over an average retest period of 2.49-years, the greatest reliability was provided by the Performance IQ

summary score (.79). The reliability coefficients for the Full-Scale IQ (.73), and Verbal IQ (.53), were both lower than comparable comparisons seen in Brown (1987). In addition, more variability was evident in the subscales, with coefficient values ranging from .27 (Similarities) to .71 for Object Assembly (Paniak, 1987).

The results of the investigation by Donaghy (1988) were very similar to those of the Brown (1987) study. A Full-Scale IQ reliability coefficient of .81 was reported. Coefficients of .76 and .77 were associated with Verbal and Performance IQ scores, respectively.

Short-term Test-retest Reliability

Studies of the test-retest reliability of the WISC over shorter periods of time were carried out by Irwin (1966), Quereshi (1968), Reger (1962), and Turner, et al. (1967).

An indication that WISC raw scores may be more stable for an older-child population than for a younger-child population was revealed in the results of the study by Irwin (1966). Two samples were used in this particular investigation. The first sample consisted of 59 young children, ranging in age between 5-years, 7-months and 6-years, 6-months (mean age = 5-years, 11-months). The second sample was comprised of the same number of older children ranging in age from 10-years, 10-months to 11-years, 7-months (mean age = 10-years, 10-months). Test-retest correlations for both samples were generated for the three

WISC summary indices, and 11 of the subtests (excluding the Mazes subtest), with a three- to four-week retest interval (Irwin, 1966).

The resulting correlation coefficients for each of the samples in Irwin (1966) were based upon test-retest comparisons of the raw scores. For all except one subtest (Picture Arrangement) the retest correlation coefficients were greater for the older-children sample. Four of the Verbal subtests (Comprehension, Arithmetic, Similarities, Vocabulary), two of the Performance subtests (Block Design, Object Assembly), and the Verbal IQ summary score, demonstrated significant differences ($p < .05$) between ages (Irwin, 1966).

A visual inspection of the resulting correlations for the samples revealed a difference in the patterns of reliabilities for each sample. The younger-children sample demonstrated somewhat greater correlation values for subtests within the Performance dimension than for those in the Verbal dimension. Accordingly, the Performance IQ summary score was also greater than the Verbal IQ score for this sample. A greater degree of similarity between verbal and performance subtests is seen in the older-children sample.

Criticism of earlier studies of the test-retest reliability of the WISC was presented by Quereshi (1968). The author stated that the findings that emerged from some

earlier studies (Coleman, 1963; Gehman & Matyas, 1956; Throne, et al., 1962; Whately & Plant, 1957) lacked generalizability beyond their particular samples because of the following: Restricted sample size, the absence of random sampling from populations, and the constrained characteristics of the samples employed (Quereshi, 1968).

With these methodological concerns in mind, Quereshi (1968) administered the WISC to 328 children, and then retested them after a three-month interval. The children ranged in age from 5- to 14-years old, and were divided into five age categories (5-, 7-, 9-, 11-, 13-years). Except for the 5-year-old age level, each child was retested within his or her original age category. Pearson-r correlations were calculated for each of the 11 subtests employed (Mazes subtest was not given), and for the three summary indices of the test (Quereshi, 1968).

Results of this study provided additional support for the results of Irwin (1966). While five of the 11 subtests for the 5-year old age category have reliability coefficients in excess of .60, and none above .70, this rate increased in the oldest age category. For the oldest age level (13-year-olds) the number of reliability coefficients greater than .60 rises to nine of 11 subtests, with five of these being greater than .70 (Quereshi, 1968). Similar to the findings of Irwin (1966), visual inspection of the results of Quereshi (1968) reveals generally greater

reliability for the Performance subtests than for the Verbal subtests in the younger age levels. This difference disappears in the older age levels.

The results of Quereshi (1968) provided an additional difference in the reliabilities for younger and older children was apparent in the summary indices of the two extreme age categories. The youngest category yielded coefficients of reliability equalling .801, .780, and .680 for Full-Scale IQ, Performance IQ, and Verbal IQ, respectively. These values are relatively lower than those for the oldest age category (Full-Scale IQ = .892; Verbal IQ = .853; Performance IQ = .823). Differences in the degree of reliability of the Verbal and Performance IQ indices for younger and older children are also apparent. In the younger age category, the Performance IQ summary score has a greater coefficient of reliability value than does the Verbal IQ score. This pattern is reversed in the older age category (Quereshi, 1968). Quereshi interprets this finding as opposing theories that predict that the Verbal summary score is intrinsicly more stable than the Performance summary score.

The results of studies that have investigated the short-term consistency of the WISC with samples drawn from the mentally retarded child population indicate a fair degree of test-retest reliability in the scores of such children.

The short-term test-retest reliability of the performance of mentally retarded children on the WISC is included in the study by Reger (1962). High test-retest reliability and stability were found in a group of 39 mentally retarded boys (Full-Scale IQ < 90) whose ages were between 11-years, 0-months, and 14-years, 11-months (Reger, 1962). Each child was re-administered the WISC between three- and four-months after the initial administration. All resultant Pearson-r coefficients in this study were greater than .67, with five of the subtests administered (Comprehension, Arithmetic, Picture Completion, Block Design, Coding) providing coefficients in the .80s. Full Scale (.95), Verbal (.92), and Performance IQs (.89) were all high (Reger, 1962). Negligible differences between the means of these three summary scores over the two intervals indicated some degree of short-term stability for these scores (Reger, 1962).

Friedman (1970) re-administered the WISC, after 17 months, to a sample of 44 children whose initial mean IQs were all below 82. The average age of this sample was 8-years, 11-months. The greatest reliability was seen in the measure of Performance IQ (.78), with Full-Scale and Verbal IQs providing coefficients of .68 and .48, respectively (Friedman, 1970).

Turner et al. (1967) employed a sample of 26 young male psychiatric patients. Employing a retest interval of six-

months, Turner et al. found a high degree of agreement for the WISC summary indices. Full-Scale and Verbal IQ provided coefficients of reliability above .80, while that for Performance IQ equalled .73 (Turner, et al., 1967). The range of subtest reliability coefficients was from .11 (Picture Arrangement) to .82 for Vocabulary (Turner, et al., 1967). The authors suggest that the retest reliability of this test is reasonable over a six-month period of time with a psychiatric population, and that the results of this study were comparable to the results of studies employing non-psychiatric samples (Turner, et al., 1967).

Test-retest Reliability of the WRAT Variables

Several sources of information are available regarding the reliability of Wide Range Achievement Test (WRAT) subtest scores over more than one administration (Brown, 1987; Donaghy, 1988; Eno & Woehlke, 1980; Naglieri & Parks, 1980; Paniak, 1987; Stevenson, Parker, Wilkinson, Hegion, & Parks, 1976; Woodward, Santa-Barbara, & Roberts, 1975).

Long-term Test-retest Reliability

The studies to be discussed (Brown, 1987; Donaghy, 1988; Eno & Woehlke, 1980; Naglieri & Parks, 1980; Paniak, 1987; Stevenson et al., 1976) examined the effect of longer retest intervals upon the test-retest reliability of the WRAT. In general, correlations not exceeding .70 have resulted; These tend to decrease as the time interval increases.

In a one-year test-retest study, Naglieri and Parks (1980) administered the WRAT to 115 children (mean age = 6-years, 10-months). The resulting raw scores were converted into two sets of standard scores based upon the norms for the 1965 and 1978 editions of the test. Twelve months later, all the children were re-administered the WRAT, and the sets of standard scores which resulted were correlated with the pretest data (Naglieri & Parks, 1980). Reliability coefficients for standard scores based upon the 1965 and 1978 versions of the WRAT were very similar. Coefficients of reliability ranged from .60 to .70 for the three subtests (Naglieri & Parks, 1980).

Some support for the Naglieri and Parks (1980) findings is seen with those from Stevenson et al. (1976). In this study the pre-kindergarten cognitive correlates of successful reading and arithmetic in elementary school were investigated. As part of this longitudinal investigation, the Reading and Arithmetic subtests of the WRAT were administered to a sample of 255 children pre-enroled in kindergarten all of whom were matched according to sex. Further assessments were conducted when these children had entered Grade One (N = 255), Grade Two (N = 142), and once again in Grade Three (N = 153). All possible test-retest comparisons were calculated separately for males and females (Stevenson et al., 1976).

The consistency of results over administrations were

generally lower for the earlier than later grade levels. Pearson-r reliability coefficient values for the Reading subtest ranged from .60 (Kindergarten vs. Grade One, either sex) to .92 (Grade One vs. Grade Two, males). One-year reliability coefficients for the Arithmetic subtest ranged from .45 (Kindergarten vs. Grade One, females) to .77 (Grade One vs. Grade Two, males). Two-year reliability of these same measure ranged from .59 (Kindergarten vs. Grade Two, males) to .79 (Grade One vs. Grade Three, females) for the Reading subtest; and, coefficients were between .41 (Kindergarten vs. Grade Two, females) and .70 (Grade One vs. Grade Three, males) for the Arithmetic subtest (Stevenson et al., 1976). Three-year reliabilities (Kindergarten vs. Grade Three) for the Reading subtest was .49 (males) and .57 (females). The Arithmetic subtest yielded reliability coefficients of .45 (males) and .43 (females) (Stevenson et al., 1976).

Eno and Woehlke (1980) employed an average retest interval of 58-months in their reliability study with 33 learning disabled and educationally handicapped children. The range of ages within this sample was from 6-years, 2-months to 11-years, 3-months. Psychometric test-retest reliability coefficients for the three subtests was .714 (Reading), .681 (Spelling), and .287 (Arithmetic). Significant changes were seen in the sample's second performance on the WRAT, with the average raw scores lower

than those from the first administration (Eno & Woehlke, 1980).

As part of her study, Brown (1987) examined the test-retest reliability and the stability of the three WRAT subtests over an average of 2.49-years. Details of the procedures employed in this study have been provided under the discussion of the test-retest reliability of neuropsychological measures with children. All three of the subtests provided Pearson-r coefficients in the .50s. These values are in keeping with those found in the Stevenson et al. (1976) examination of the three-year reliability of these subtests.

Examination of the ICC for these same measures yielded a good degree of stability with coefficients ranging in the .50s. Intraclass correlation coefficient values are provided in Appendix E.

Donaghy (1988) reports Pearson-r correlation coefficients ranging from .56 (WRAT Reading) to .50 (WRAT Arithmetic) for his combined sample of children.

Slightly dissimilar findings of psychometric test-retest reliability are provided by Paniak (1987). The author used an average retest interval of 2.65-years with his heterogeneous clinical sample of 75 children. Resulting Pearson-r correlations coefficients are .52 (Reading), .42 (Spelling), and .51 for the Arithmetic subtest (Paniak, 1987), slightly lower than those found in the Brown (1987)

and Donaghy (1988) studies.

Short-term Test-retest Reliability

The short-term reliability of this test is generally adequate, with Pearson-r correlation coefficients of reliability in the .80s commonly reported in the literature (Woodward et al., 1975). Jastak and Jastak (1978) suggest that the three-month test-retest reliability of the WRAT is adequate.

A sample of emotionally disturbed children (N = 43), with a mean age of 10.6-years, was administered the WRAT twice within an average of 155 days (Woodward et al., 1975). Pearson-r reliability coefficients for the three WRAT subtests were all greater than .85 (Woodward et al., 1975). Woodward, et al. (1975) report that practice effects had occurred in only the Reading subtest. The authors attempted to replicate these results by examining the reliability of the WRAT, administered in test-retest fashion within 15 days, to a second sample of 63 youths. The average age of this second sample was 10.6-years, and, all children were previously diagnosed as being emotionally disturbed (Woodward et al., 1975). The pattern of the reliability coefficients is reported to have closely resembled that of the first sample, indicating acceptable reliability in subtest scores over short periods of time (Woodward et al., 1975).

Test-retest Reliability of the PPVT Variable

The Peabody Picture Vocabulary Test (Dunn, 1965) has been the subject of much research into its psychometric reliability. Summaries of past studies (Bochner, 1978; Dunn & Dunn, 1981), as well as more recent studies (Brown, 1987; Dean, 1980; Donaghy, 1988; Paniak, 1987) have clarified the test-retest reliability of this measure.

Long-term Test-retest Reliability

Summaries of studies concerning the long-term stability of the PPVT (Bochner, 1978; Dunn & Dunn, 1981) indicate acceptable median correlation coefficients for retest periods in excess of one year. Bochner (1978) found a median long-term test-retest reliability coefficient of .77 for five studies of the PPVT drawn from the literature existing between the years 1965 and 1972. Bochner (1978) based her reliability coefficients primarily upon the correlation of PPVT raw scores. It is worthwhile to note that each of the five studies included in this review had employed samples of mentally retarded or disadvantaged children (Bochner, 1978). Bochner (1978) reports that the literature suggests that PPVT raw scores appear to be the most stable for the mentally retarded samples studied.

Dunn and Dunn (1981) provide additional information concerning the long-term stability of this instrument. The authors present median correlation coefficient values for PPVT-IQ scores based upon a number of studies providing reliability data on this test between the years 1965 and

1979 (Dunn & Dunn, 1981). For retest intervals greater than one year, Dunn and Dunn (1981) report a median test-retest reliability coefficient of .59 for PPVT-IQ scores.

The Dunn and Dunn (1981) study concurred with Bochner (1978) on the relative stability of PPVT scores in mentally retarded or institutionalized populations. The least stable coefficients of reliability were reported to be exhibited by disadvantaged children (Dunn & Dunn, 1981).

The findings of the Brown (1987), Paniak (1987), and Donaghy (1988) studies concerning the long-term reliability of a number of tests and measures with a heterogeneous clinical sample of children also included test-retest reliability data on the PPVT-IQ. The findings of these studies are in complete agreement with the median value presented by Dunn and Dunn (1981). The Pearson-*r* correlation coefficient values for PPVT-IQ scores in the Brown (1987) and Paniak (1987) studies are provided in Appendix E.

The study by Brown (1987) included an estimate of the long-term stability of PPVT-IQ scores within individuals. The resultant intraclass correlation coefficient yielded by her analysis was identical (.71) to the Pearson-*r* test-retest correlation coefficient (Brown, 1987). The PPVT-IQ scores of the clinical sample used in Brown (1987) would appear to have excellent stability as well as consistency over the course of that particular study.

Short-term Test-retest Reliability

Studies of the test-retest reliability of this instrument over retest intervals less than one year (Bochner, 1978; Dean, 1980; Dunn & Dunn, 1981) have indicated higher reliability coefficients than those occurring for longer intervals. Shorter intervals are expected to result in increased reliability for the PPVT scores (Dunn & Dunn, 1981) as they would for any test measure (Anastasi, 1982).

Bochner (1978) examined the findings of previous studies on this topic and calculated a median coefficient of reliability equal to .74 for PPVT raw scores. Dunn and Dunn (1981) provided nearly identical findings to those of Bochner (1978). In their report, the authors reported a median correlation coefficient value of .72 for PPVT-IQ scores (Dunn & Dunn, 1981). The authors advance the expectation that PPVT-IQ scores should demonstrate slightly less stability than PPVT raw scores (Dunn & Dunn, 1981).

Dean (1980) provided findings not included in the Bochner (1978) or the Dunn and Dunn (1981) reviews on the test-retest reliability of the PPVT with emotionally disturbed children. Fifty-two children (mean age = 15.17-years) were administered the PPVT twice, with a retest interval of approximately six months. The results of Dean (1980) provided a Pearson-r correlation coefficient of .57 for PPVT raw scores. Apparently, a lesser degree of

stability for PPVT raw scores may be evident in emotionally disturbed populations.

Test-retest Reliability of the MAT Variables

Only one study (Zingale, Smith, & Dokecki, 1980) was found that directly assessed the test-retest reliability of the Metropolitan Achievement Test (MAT). This study examined the temporal stability of the three levels of the MAT (Primary I, Primary II, and Elementary) by comparing the one-month test-retest reliability coefficients of the MAT subtests against a standard of .80 (Zingale et al., 1980). Only Form-H of the MAT was employed at any time.

Of the verbal subtests (Word Knowledge, Reading, Total Reading), Total Reading exceeded the .80 criterion for all three MAT levels. The highest reliability for Total Reading was seen in the Elementary level (.97). Only the Word Knowledge and Reading subtests of the Elementary level exceeded a reliability coefficient of .80, providing values of .94 and .92, respectively.

The math subtests (Math Computation, Math Concepts, Math Problem Solving, and Total Math) were found to be somewhat less stable in this study. The reliability of scores from the Total Math subtest was highest for the Primary I battery (.90), and less so in the Primary II (.84) and Elementary levels (.79). None of the remaining math subtests achieved retest correlation coefficients exceeding .80 for their respective MAT levels; however, Zingale, et

al. (1980) report that most were in the .70s.

Purpose of Study

Despite the existence of a large quantity of literature within the field of neuropsychological assessment, comparatively little research has been focused upon the consistency and stability of neuropsychological test scores over time. In particular, these psychometric concerns have not attracted the attention of those clinicians who employ neuropsychological tests with child populations.

Three recent studies have emerged that have addressed the issue of the consistency of children's scores on 50 neuropsychological and psychological measures over time (Brown, 1987; Donaghy, 1988; Paniak, 1987). Estimations of the internal stability of these scores were also included in the Brown (1987) study through the use of the ICC. All three researchers studied heterogeneous clinical samples of children; their results contributing to the information available concerning the test-retest reliability and stability of the particular measures investigated when used with clinical populations of the type typically assessed at a large assessment center.

In addition to answering important questions concerning these psychometric issues with a heterogeneous clinical population of children, the studies by Brown (1987), Paniak (1987), and Donaghy (1988) pose additional questions for future research. Each study stressed the need for more

research examining the retest reliability of neuropsychological measures using different representative diagnostic categories of subjects (Brown, 1987; Paniak, 1987). Through such studies, the variability of test scores over time can be determined for specific patient populations (e.g., children with specific learning disabilities).

Recommendations for investigations of this type would also have to consider some of the major limitations of the Brown (1987), Paniak (1987) and Donaghy (1988) studies. One such limitation is the large variability in retest intervals within each of the Brown (1987) and Paniak (1987) studies. Brown (1987) reports that the range of retest interval in her retrospective study was from as little as one month, to as great as 12 years. Similar variability in retest intervals is reported in Paniak (1987). This limitation also necessarily affected the study by Donaghy (1988). A limitation to generalization reported by two of the authors (Brown, 1987; Paniak, 1987) was the inclusion of only children who were referred for repeated clinical assessment. Brown (1987) comments that this sub-population of children continued to exhibit behavioral or academic difficulties of a serious enough nature to recommend re-evaluation. The results of the study, however, remain particularly applicable to such children.

Finally, the use of the test-retest paradigm for studying neurospychological measures has been criticized, as

it invites contamination of results through practice and memory effects (Shaw, 1966; Klonoff et al., 1971). However, as Klonoff et al., (1971) clearly state, "clinical practice often dictates the readministration of psychological tests" (p. 292). Given this reality of clinical practice, the clinical neuropsychologist must be aware of the test-retest reliability of tests that he or she may repeatedly administer to patients. Only with such information can the clinician guage his or her confidence in the validity of changes observed in a patient's test performance over time (Rourke et al., 1986).

The current research consists of four investigations into the test-retest reliability and consistency of neuropsychological measures when used with children.

Investigation 1

The concern, arising from earlier investigations (Brown, 1987; Donaghy, 1988; Paniak, 1987), that test-retest reliability studies employ representative diagnostic categories of subjects, and that the variability within retest intervals be minimized, are addressed by this first investigation. The purpose of this investigation is to examine the two-year retest reliability and consistency of 64 neuropsychological measures (listed in Appendix F) for a homogeneous sample of disabled readers. For the purposes of this study, the disabled readers are homogeneous only to the extent that they share difficulties in reading ability, as

assessed by the MAT. No attempt is made to equate subjects on the basis of the qualitative or quantitative aspects related to their reading disability.

Investigation 2

The second investigation seeks to determine if the retest reliability of the neuropsychological measures differ for the disabled readers sampled as compared to a sample of normal readers. The two-year retest reliability and consistency of the same 64 measures was calculated for a sample of normal readers that were matched for grade with the disabled readers. The pattern of reliability coefficients was compared between the normal readers and disabled readers samples.

Investigation 3

The pattern of correlation coefficients generated by the disabled readers sample will be compared to the correlation coefficients in the Donaghy (1988) study and the Brown (1987) study in the third investigation. The purpose of this investigation is to determine if the sample of disabled readers differs from a heterogeneous clinical sample in terms of the test-retest reliability or consistency of the neuropsychological measures that are common to both studies.

Investigation 4

This final investigation will provide a partial clinical validation of the pattern of correlations arising

from the analysis of the combined disabled and normal readers groups. The two-year test-retest reliability and consistency of this combined group at Year-2 and also at Year-4 will be calculated. This resulting ranking of coefficients for measures common to all ages will then be compared to the pattern of coefficients calculated for the combined group at Year-0 and Year-2. Direct validation of the normal readers or disabled readers samples was precluded because of the reduced sample size over four years.

CHAPTER II

METHOD

Subjects

The data used in this study were collected during the course of a previous longitudinal investigation by Rourke, Ridgley, & Orr (1973). In that study, a group of normal male readers was compared to a group of male disabled readers on a series of neuropsychological tests over two-year and four-year periods of time.

At the time of the Year-2 assessment, 27 children, ranging in age from 109-months to 122-months (mean age = 115.28-months) comprised the Normal-Readers (NR) group. Selection of children into this group was made on the basis of their MAT subtest scores. Children whose centile score on the Reading subtest of the MAT, was 50 or greater and whose score on either the Word Knowledge or Word Discrimination subtests was 60 or above were included into the NR group.

The subjects in the Disabled-Readers (DR) group all had a centile score of 20 or lower on the Reading subtest of the MAT. In addition, these children earned centile scores of 35 or lower on the MAT Word Knowledge or the Word Discrimination subtests. The 25 children in the DR group

had an age range of between 109-to 122-months, and a mean age of 115.28-months. Subjects in both groups had WISC Full-Scale IQ scores in excess of 80. Specifically, Full-Scale IQ scores ranged from 95 to 132 for the NR group, and from 81 to 125 for the DR group.

Twenty-three children comprised the NR group after the assessment at Year-4, while the DR group contained 19 children.

Test Measures

A list of the tests and measures employed in this study is included in Appendix F. Only those tests that were administered in the same manner to the subjects in the first and second assessment periods were included in this study. The tests presented in Appendix F represent the same content and manner of administration for each child, and at each age level (7 to 8 years and 9 to 10 years).

The tests were categorized into the ability domains that they are thought to assess. The eight ability domains chosen for the present study are: (1) Psychometric Intelligence; (2) Academic Achievement and Reading; (3) Auditory-Perception and Language; (4) Visual-Perception; (5) Tactile-Perception; (6) Motor; (7) Right-Left Awareness; and, (8) Underlining Test measures. The particular categorical organization employed in this study represents a slight modification of the schema used in Paniak (1987), in order to allow incorporation of the same categories utilized

by Rourke et al. (1973). Greater discussion and explanation of what each test seeks to assess is available to the reader in Rourke, Fisk, and Strang (1986), and Reitan and Davison (1974).

Procedure

The procedures followed in administering the tests to the children are those that were employed in Rourke et al. (1973). A brief summary of these procedures may prove useful to the reader, and will be provided here.

At the initiation of the study, the MAT was administered to a number of male students at seven public elementary schools in the Ontario region. The WISC was administered only to those children who met the MAT selection criteria. The children comprising the NR group earned centile scores of 50 or greater on the MAT Reading subtest. Subjects in the two groups were paired for grade level, and each subject was administered the battery of tests individually. During the second administration of tests, the children participating in the first segment of this study were readministered the MAT in a group format, following which individual testing was initiated. At each testing situation every effort was made to motivate the children to perform to their best ability.

The number of psychometrists administering the tests varied from one to three individuals for each school, and the battery of tests was presented in a more or less random

basis. Each psychometrist was blind to the MAT scores of the children who were assessed, and each tested roughly the same number of children from the two groups.

Statistical comparisons to be utilized in the present investigation to analyse the data employed the SPSSx statistical computer program. Four investigations were conducted in this study.

Investigation 1

Pearson product-moment correlation coefficients were computed between DR subjects' initial and second assessment scores on each test measure. The resultant coefficients of reliability were interpreted as estimations of the consistency of these measures over the two year period. Comparisons for all the measures were then repeated, with the intraclass correlation coefficient serving as the estimate of reliability. The internal stability of the test measures was interpreted from the resultant correlation coefficients.

Investigation 2

Pearson product moment correlation coefficients, and ICC coefficients, were computed between NR subjects' initial and second assessment scores on each test measure. A Spearman rank-order correlation was performed in order to compare the coefficients of reliability generated for the NR and DR groups. The result of this comparison was provided in the Spearman rank-order correlation coefficient.

Investigation 3

The Pearson-r coefficients generated for the DR group as part of Investigation 1 were compared to the Pearson-r coefficients that resulted from the Donaghy (1988) study. Only those tests that were administered in the same manner to the subjects of both studies were compared. The comparison consisted of the calculation of a Spearman rank - order correlation. A second analysis was undertaken examining the ICC coefficients generated for the DR subjects, and the results of the Brown (1987) study. The same restriction upon the administration of tests was utilized to determine which measures were to be compared. Again, the comparison consisted of a Spearman rank-order correlation.

Investigation 4

Pearson product moment correlation coefficients were calculated for all subjects pooled together and assessed at Year-2 and again at Year-4. Similarly, ICC coefficients were generated for these same subjects over the same two-year periods. The combination of DR and NR subjects into a single pooled group was necessary in order to preserve the sample size. Ten children (4 from NR and 6 from DR groups) were dropped from this portion of the investigation due to missing data. As a result the number of children comprising this pooled group was 42. The loss data associated with the 10 children dropped from the study was assumed to be due to

normal attrition and not due to any systematic factor. The results of these calculations were compared to the results of similar pooled-groups calculations over the initial two years (Year-0 and Year-2). Only those tests administered in the same manner to all subjects were examined. A Spearman rank-order correlation was calculated for the two types of coefficients.

CHAPTER III

RESULTS

The results of the current study will be presented separately for each of the four investigations previously outlined. These investigations focus upon the consistency and stability of the NR and DR childrens' performances on a variety of neuropsychological measures over a two-year retest interval.

The Pearson-r coefficient served as the index of consistency for the analyses used in the present study. All Pearson-r coefficients were calculated through the use of the PEARSON CORR computer program of the SPSSx (Nie, 1983) statistical batch system.

The retest stability of a measure is the extent to which confounding variables do not impact upon the relationship of that measure's initial assessment value with subsequent assessment values. The ICC coefficient served as the measure of retest stability in the current study.

Based upon an analysis of variance (ANOVA) model, the formula for the calculation of the ICC is given by Berk (1979):

$$ICC = \frac{BMS - EMS}{BMS + (k-1)\ EMS}$$

where BMS represents the mean square value for childrens'performance, EMS represents the residual mean square value, and k equals the number of assessments performed.

The ICC is sensitive to the stability of childrens' test performance over time because significant differences between test score means and variances will attenuate the value of the ICC. The ANOVA components for each neuropsychological measure examined were calculated using the RELIABILITY computer program of the SPSSx (Nie, 1983) batch system. The model employed was a 1 X 2 mixed effects design, with childrens' performance a random factor having one level and assessment periods a fixed factor having two levels. The significance level for each ICC coefficient was arrived at by examining the F-statistic associated with the factor of interest.

Investigation 1

This first investigation was concerned with the degree of consistency and stability of the 25 DR childrens' performance on a variety of neuropsychological measures over a two-year period of time. The means and standard deviations of the childrens' performances at the initial, and subsequent assessment periods, are provided in Table 1. The neuropsychological variables, grouped by ability domain,

Table 1

Descriptive Statistics of Test Variables for Disabled Readers

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Psychometric Intelligence				
Full Scale IQ	101.1	6.4	100.8	10.1
Verbal IQ	98.0	9.2	97.6	8.5
Performance IQ	104.9	9.6	104.4	12.6
Information Scaled Score	8.2	2.3	8.4	1.7
Comprehension Scaled Score	10.9	3.0	9.5	2.8
Digit Span Scaled Score	9.1	2.4	8.4	3.0
Arithmetic Scaled Score	10.1	2.4	9.8	2.0
Similarities Scaled Score	10.0	3.1	10.4	1.7
Vocabulary Scaled Score	9.5	2.1	10.4	1.7
Picture Completion Scaled Score	12.1	2.4	10.1	2.6
Picture Arrangement Scaled Score	11.0	2.2	10.7	2.8
Block Design Scaled Score	10.7	2.9	10.7	2.7
Object Assembly Scaled Score	10.6	3.0	10.7	2.7
Coding Scaled Score	9.2	2.8	10.5	2.2

Table 1 (Continued)

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Academic Achievement and Reading				
MAT Word Knowledge Standard Score	35.7	5.4	44.5	9.0
MAT Word Discrimination Standard Score	40.1	7.2	44.2	8.6
MAT Reading Standard Score	32.3	5.5	39.5	8.3
WRAT Reading Standard Score	92.0	7.2	95.1	9.7
WRAT Spelling Standard Score	93.3	7.8	95.1	9.7
WRAT Arithmetic Standard Score	95.1	6.3	92.2	6.1
Auditory-Perception and Language				
PPVT-IQ	104.0	12.8	103.9	11.8
Auditory Closure correct	9.7	4.2	13.9	3.6
Sentence Memory correct	10.5	2.1	12.4	2.1
Verbal Fluency correct	4.1	2.0	7.1	2.2
Auditory (R) errors	.1	.2	-- *	-- *
Auditory (L) errors	.1	.2	.1	.2
Speech-sounds Perception correct	16.9	6.3	23.5	3.5

Table 1 (Continued)

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Motor				
Finger Tapping (R) taps/10"	28.1	5.5	32.4	5.0
Finger Tapping (L) taps/10"	26.0	5.3	31.0	4.5
Foot Tapping (R) taps/10"	25.2	5.2	28.1	3.2
Foot Tapping (L) taps/10"	23.2	4.2	26.0	3.1
Maze Speed (R)	109.4	28.4	108.3	20.7
Maze Speed (L)	113.4	35.0	103.6	23.4
Maze Time (R)	5.4	4.2	2.4	1.5
Maze Time (L)	10.5	5.8	6.6	3.4
Maze Counter (R) errors	36.4	20.6	19.1	12.4
Maze Counter (L) errors	68.9	32.9	45.8	20.9
Grip Strength (R) Kg.	12.3	2.1	13.2	2.9
Grip Strength (L) Kg.	11.7	2.2	12.2	2.9
Name Writing speed (R)	18.4	6.6	13.9	6.5
Name Writing speed (L)	32.9	14.1	27.1	10.8
Visual-Perception				
Target correct	12.6	2.8	15.6	2.7
Rhymes correct	12.9	2.8	15.6	2.7

Table 1 (Continued)

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Reverses correct	31.3	4.2	38.1	5.2
Visual (R) errors	.2	.5	.2	.5
Visual (L) errors	.4	.6	.1	.4
Tactile-Perception				
Finger Agnosia (R) errors	2.0	1.9	.5	1.1
Finger Agnosia (L) errors	2.7	2.4	.7	1.0
Tactile (R) errors	.9	1.2	.4	.8
Tactile (L) errors	.7	1.1	.1	.4
Underlining Test				
Subtest 1	24.6	7.9	23.8	5.8
Subtest 2	25.8	7.3	24.7	5.0
Subtest 3	15.4	5.1	14.0	4.4
Subtest 4	6.1	4.9	6.8	4.4
Subtest 5	19.2	5.9	18.9	3.8
Subtest 6	13.7	5.4	13.6	3.1
Subtest 7	14.4	5.0	12.7	3.1
Subtest 8	7.3	3.1	6.6	1.1
Subtest 9	7.3	3.0	6.0	2.2
Subtest 10	8.5	3.6	9.5	4.1

Table 1 (Continued)

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Subtest 11	7.3	4.0	11.9	4.4
Subtest 12	5.2	2.6	5.9	1.4
Subtest 13	23.1	8.6	19.2	5.2
Right-Left Awareness				
Right-Left Awareness correct	16.0	3.6	17.5	4.0

are presented in Appendix F.

Presenting the results concerning the consistency of the childrens' performance first, the range of Pearson-*r* values within each of the eight ability realms were examined as well as how these values ranked against the other neuropsychological measures in terms of consistency. Table 2 contains the Pearson-*r* values, ranked in terms of high to low value associated with the DR childrens' performance. These same coefficient values are presented within an organization of the ability areas with which they are associated in Table 3.

A large amount of variation in the Pearson-*r* coefficient values was in evidence for the DR sample's assessment over the two time periods. The greatest degree of consistency was found for the Rhymes subtest of the Children's Word-Finding Test (Rhymes correct), which had a coefficient of .73. In contrast, the least consistent performance was observed for Subtest 9 of the Underlining test ($r = -.39$). The mean Pearson-*r* coefficient calculated for these measures was found to be .24 ($SD = .26$). Calculation of a correlation coefficient for the Auditory (R) errors measure was precluded by insufficient data. Only 21 of the 63 resulting Pearson-*r* coefficients were significant to the $p < .05$ level. The consistency of the measures within each ability domain are examined next.

Table 2

Rank Order of Disabled Readers' Tests Based on Magnitude of Correlation (Pearson-r)

Rank	Variable	Pearson-r
1	Rhymes correct	.73
2	Target correct	.65
3	Sentence Memory correct	.63
4	Grip Strength (R) Kg.	.63
5	Grip Strength (L) Kg.	.60
6	PPVT-IQ	.57
7	Finger Tapping (L) taps/10"	.56
8	Full Scale IQ	.56
9	Maze Speed (L)	.55
10	Performance IQ	.50
11	Foot Tapping (R) taps/10"	.48
12	Verbal IQ	.48
13	MAT Word Discrimination Standard Score	.47
14	WRAT Arithmetic Scaled Score	.46
15	Finger Tapping (R) taps/10"	.46
16	Maze Counter (R) errors	.46
17	Auditory Closure correct	.44
18	Information Scaled Score	.43
19	Maze Time (L)	.41
20	Block Design Scale Score	.39
21	Maze Counter (L) errors	.38
22	Foot Tapping (L) taps/10"	.36*
23	MAT Reading Standard Score	.34*
24	Similarities Scaled Score	.34*
25	Digit Span Scaled Score	.33*
26	Tactile (L) errors	.33*
27	WRAT Spelling Scaled Score	.31*

Table 2 (Continued)

Rank	Variable	Pearson-r
28	WRAT Reading Scaled Score	.31*
29	Maze Time (R)	.29*
30	Speech-sounds Perception correct	.28*
31	Coding Scaled Score	.28*
32	Name Writing speed (R)	.27*
33	Underlining Subtest	.27*
34	Picture Arrangement Scaled Score	.27*
35	Object Assembly Scaled Score	.25*
36	Reverse correct	.25*
37	Maze Speed (R)	.24*
38	Vocabulary Scaled Score	.24*
39	MAT Word Knowledge Standard Score	.24*
40	Underlining Subtest 4	.22*
41	Underlining Subtest 11	.20*
42	Comprehension Scaled Score	.18*
43	Picture Completion Scaled Score	.13*
44	Arithmetic Scaled Score	.06*
45	Verbal Fluency correct	.05*
46	Underlining Subtest 3	.02*
47	Name Writing speed (L)	.01*
48	Finger Agnosia (R) errors	-.001*
49	Underlining Subtest 7	-.001*
50	Finger Agnosia (L) errors	-.001*
51	Tactile (R) errors	-.02*
52	Underlining Subtest 6	-.02*
53	Auditory (L) errors	-.04*
54	Underlining Subtest 12	-.05*
55	Underlining Subtest 10	-.07*
56	Underlining Subtest 5	-.09*

Table 2 (Continued)

Rank	Variable	Pearson-*r*
57	Underlining Subtest 1	-.11*
58	Right-Left Awareness correct	-.13*
59	Visual (R) errors	-.16*
60	Underlining Subtest 8	-.19*
61	Visual (L) errors	-.25*
62	Underlining Subtest 2	-.27*
63	Underlining Subtest 9	-.39*
64	Auditory (R) errors	-- ++

* $p > .05$

++ Insufficient data to calculate Pearson-*r* coefficient

Table 3

Pearson-r's for Disabled Readers' Tests Within Ability Domains

Test Measure	Pearson-r	
Psychometric Intelligence		
Full Scale IQ	.56	Good
Performance IQ	.50	
Verbal IQ	.48	
Information Scaled Score	.43	
Block Design Scale Score	.39	Fair
Similarities Scaled Score	.34	
Digit Span Scaled Score	.33	
Coding Scaled Score	.28	Poor
Picture Arrangement Scaled Score	.27	
Object Assembly Scaled Score	.25	
Vocabulary Scaled Score	.24	
Comprehension Scaled Score	.18	
Picture Completion Scaled Score	.13	
Arithmetic Scaled Score	.06	
Academic Achievement and Reading		
MAT Word Discrimination Standard Score	.47	Good
WRAT Arithmetic Scaled Score	.46	
MAT Reading Standard Score	.34	Fair
WRAT Spelling Scaled Score	.31	
WRAT Reading Scaled Score	.31	
MAT Word Knowledge Standard Score	.24	Poor
Auditory-Perception and Language		
Sentence Memory correct	.63	Excellent
PPVT-IQ	.57	Fair
Auditory Closure correct	.44	

Table 3 (Continued)

Test Measure	Pearson-r	
Speech-sounds Perception correct	.28	Poor
Verbal Fluency correct	.05	
Auditory (L) errors	-.04	
Auditory (R) errors	-- ++	
Motor		
Grip Strength (R) Kg.	.63	Excellent
Grip Strength (L) Kg.	.60	
Finger Tapping (L) taps/10"	.56	Good
Maze Speed (L)	.55	
Foot Tapping (R) taps/10"	.48	
Finger Tapping (R) taps/10"	.46	
Maze Counter (R) errors	.46	
Maze Time (L)	.41	
Maze Counter (L) errors	.38	Fair
Foot Tapping (L) taps10"	.36	
Maze Time (R)	.29	Poor
Name Writing speed (R)	.27	
Maze Speed (R)	.24	
Name Writing speed (L)	.01	
Visual-Perception		
Rhymes correct	.73	Excellent
Target correct	.65	
Reverse correct	.25	Poor
Visual (R) errors	-.16	
Visual (L) errors	-.25	
Tactile-Perception		
Tactile (L) errors	.33	Fair

Table 3 (Continued)

Test Measure	Pearson-r	
Finger Agnosia (R) errors	-.001	Poor
Finger Agnosia (L) errors	-.001	
Tactile (R) errors	-.02	
Underlining Test		
Underlining Subtest 13	.27	Poor
Underlining Subtest 4	.22	
Underlining Subtest 11	.20	
Underlining Subtest 3	.02	
Underlining Subtest 7	-.001	
Underlining Subtest 6	-.02	
Underlining Subtest 12	-.05	
Underlining Subtest 10	-.07	
Underlining Subtest 5	-.09	
Underlining Subtest 1	-.11	
Underlining Subtest 8	-.19	
Underlining Subtest 2	-.27	
Underlining Subtest 9	-.39	
Right-Left Awareness		
Right-Left Awareness correct	-.13	Poor

++ Insufficient data to calculate Pearson-r coefficient

Nine of the 14 measures of Motor ability were associated with significant correlation coefficients. The consistency of these measures ranged from .63 for Grip Strength (R) to .38 for Maze Counter (L) errors. The rankings of these measures in terms of their consistency was from 4th (Grip Strength (R)) to 47th (Name Writing Speed(L)). The majority of these measures were ranked within the upper half of the rank order distribution of consistency. The median ranking for all the Motor variables was 17. Figure 1 presents the Pearson-r coefficients for these measures within a graphical format. Descriptions of the abbreviations found in Figure 1, and in the additional figures, are provided in Appendix F.

Three measures of Auditory-Perception and Language were associated with significant Pearson-r coefficients. Sentence Memory correct (r = .63) provided the highest coefficient magnitude. Graphical representation of all Auditory-Perceptual and Language measures is presented in Figure 2. For consistency, this variable, and PPVT-IQ (r = .57) were both ranked within the top 10 measures. Two additional measures of Auditory-Perception and Language, Verbal Fluency correct (r = .05) and Auditory (L) errors (r = -.04) were ranked within the lower one-third of the rankings. For all of the Auditory-Perception and Language measures, the median consistency ranking was 24.

Only two variables within the domain of Academic

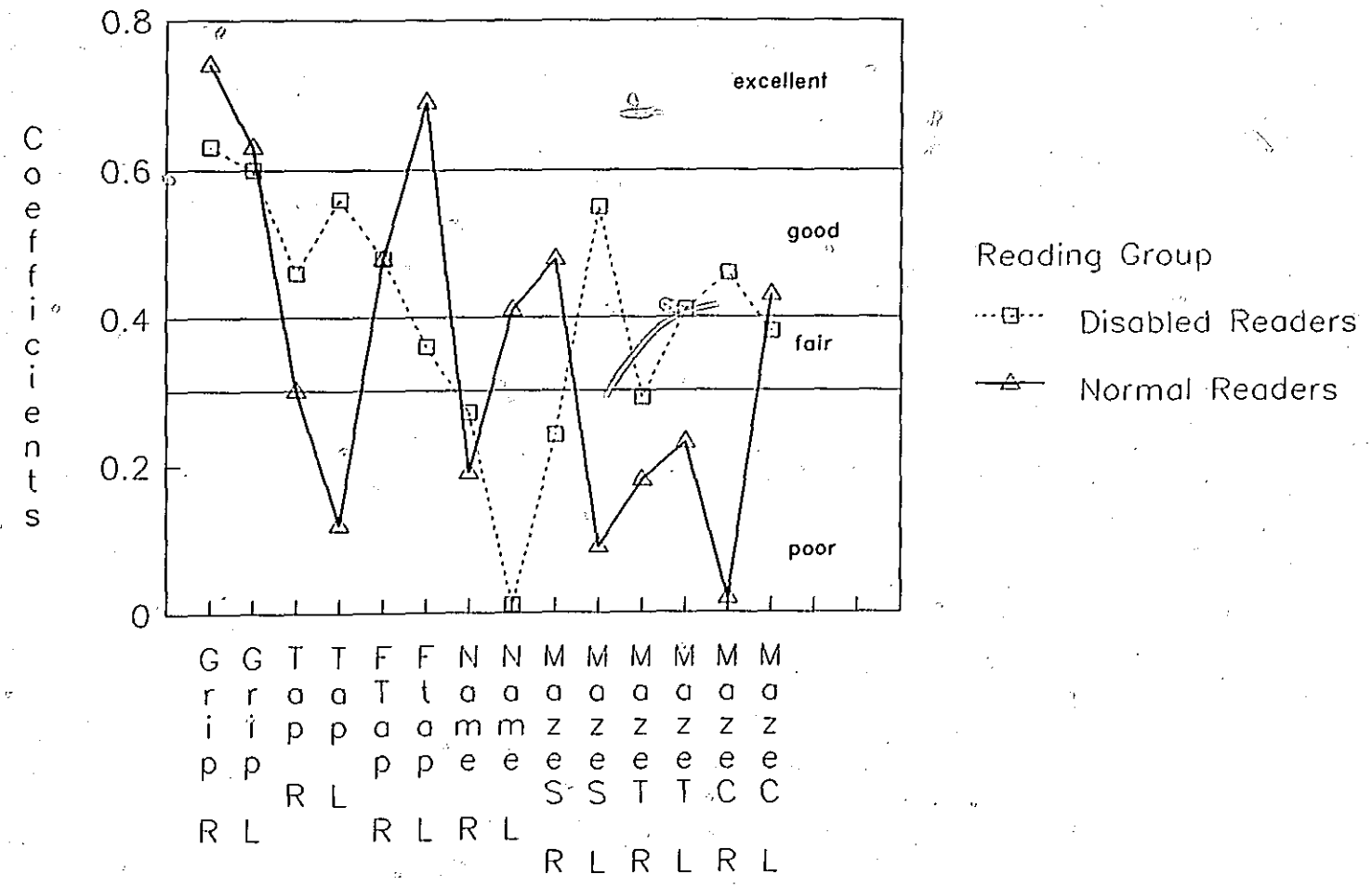

Figure 1. Consistency of motor measures. Disabled vs. Normal readers.

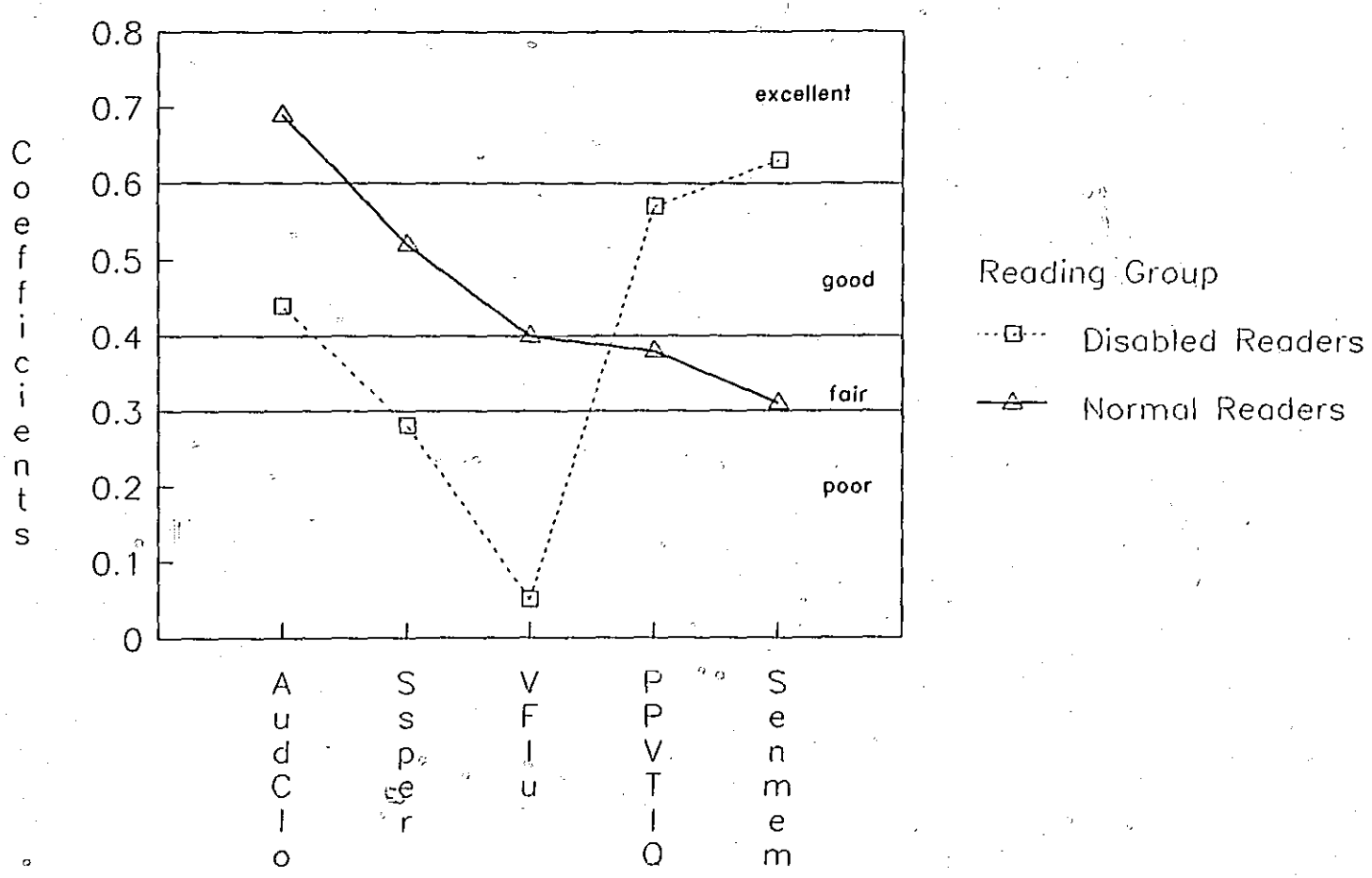

Figure 2. Consistency of auditory–perception and language measures Disabled vs. Normal readers.

Achievement and Reading provided significant Pearson-r correlation coefficients. Figure 3 presents these results within a graphical format. These measures, the MAT Word Discrimination subtest (r = .47) and the WRAT Arithmetic subtest (r = .46) were ranked 13th and 14th, respectively, in terms of their consistency. The remaining variables within this domain were ranked lower than 27. For all of the Academic Achievement and Reading measures taken as a whole, the median rank was calculated to be 33.

Examining the measures of Psychometric Intelligence, significant correlation coefficients were found for five of the 14 variables. The range of consistency for these significant measures was from .56 (WISC Full Scale IQ) to .39 (Block Design subtest); these coefficients are graphically represented in Figure 4. In contrast to the Motor variables, these measures were roughly evenly distributed through out the rank-order distribution. The median rank for the Psychometric Intelligence ability domain was 28.

The two measures of Visual-Perceptual ability that provided significant correlation coefficients (Rhymes correct and Target correct) were also the two most consistent measures overall, being associated withcoefficients of .73 and .65, respectively. The remaining measures within this domain had consistency rankings in the lower third of the rank order distribution.

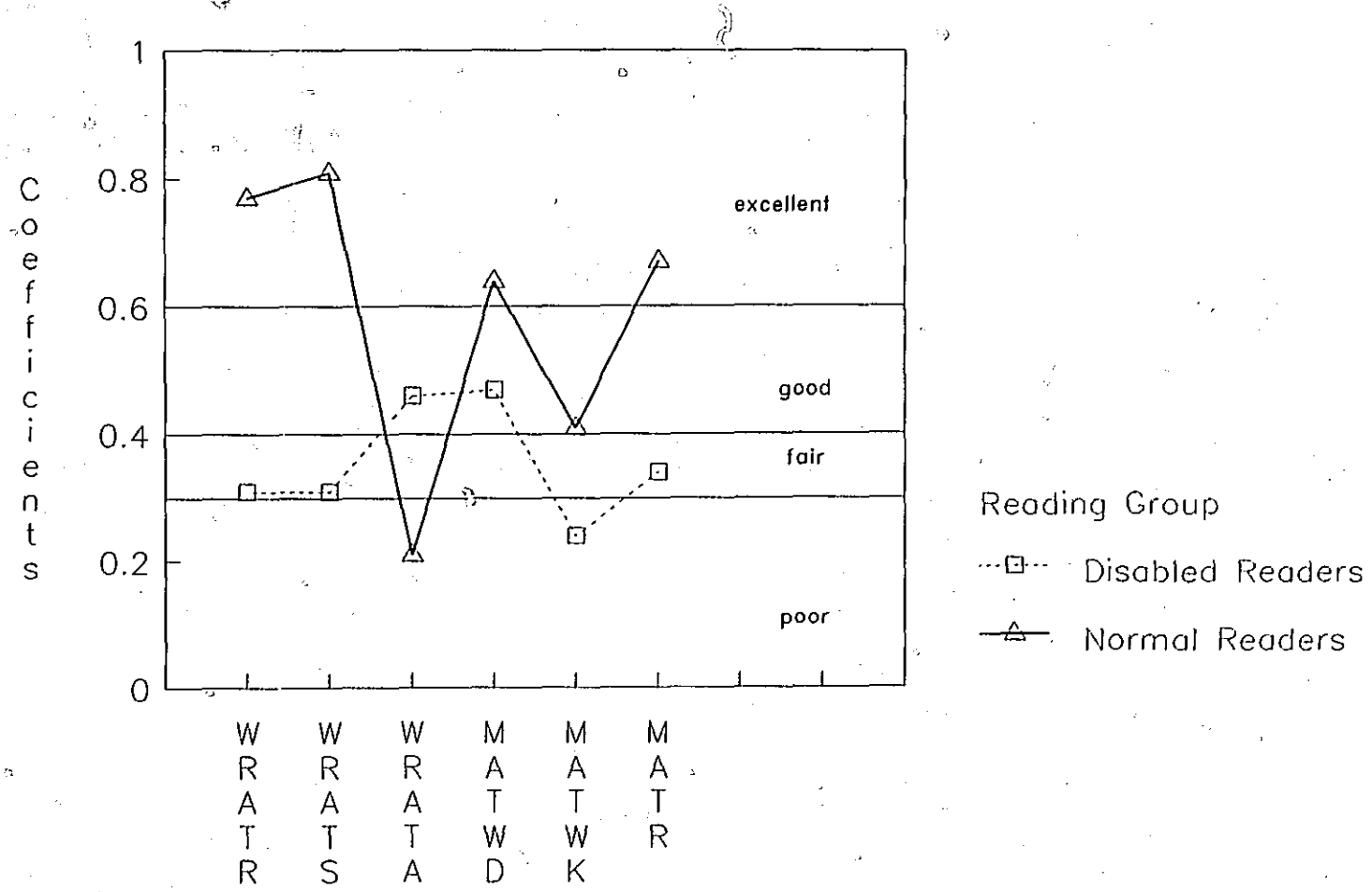

Figure 3. Consistency of academic achievement and reading measures. Disabled vs. normal readers

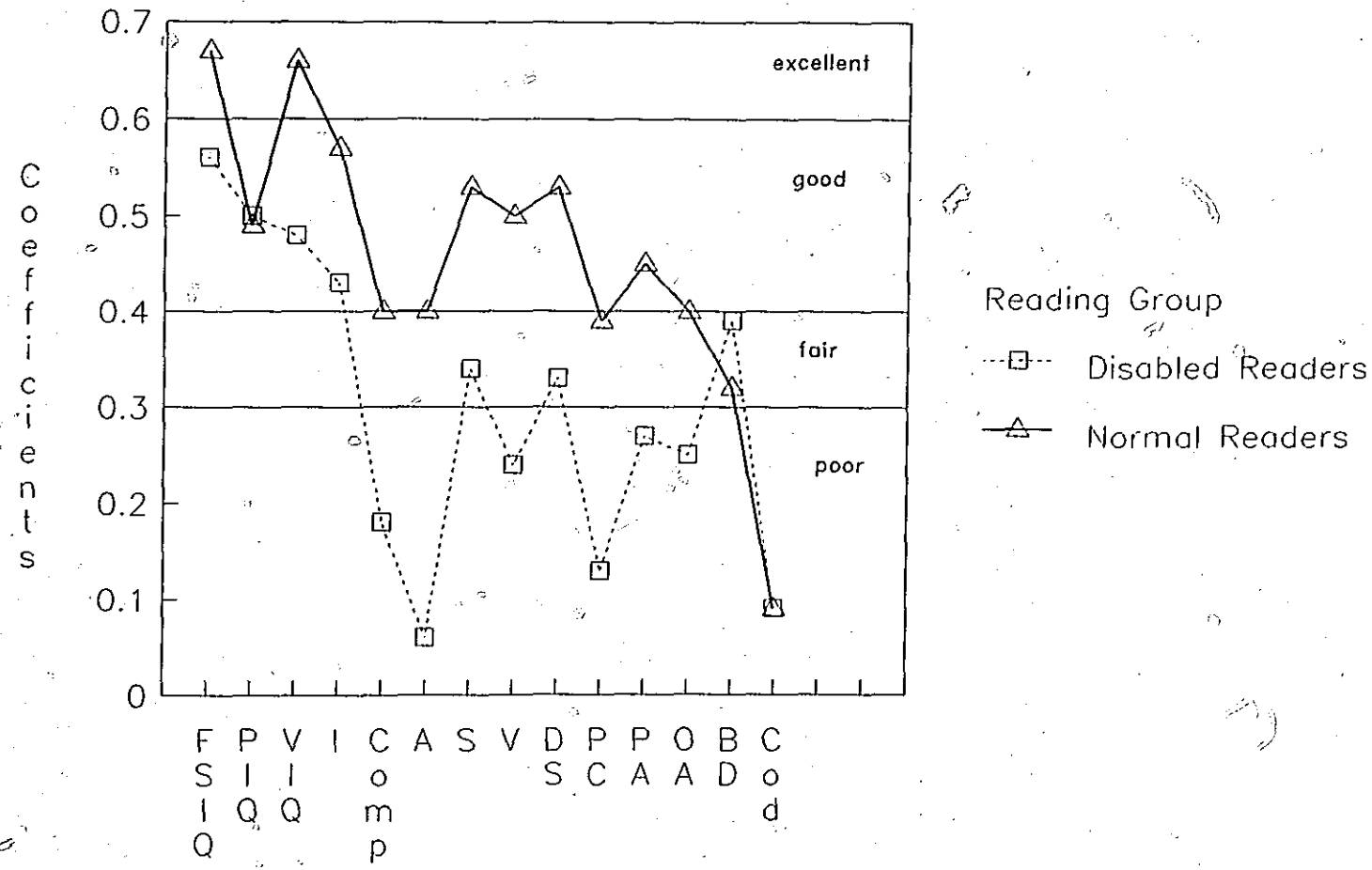

Figure 4. Consistency of psychometric intelligence measures. Disabled vs. Normal readers.

Visual (R) errors and Visual (L) errors yielded coefficients of -.16 and -.25, respectively. A graph of the coefficients associated with all of the Visual-Perceptual measures is provided in Figure 5. The median ranking for these measures was found to be 36.

None of the 13 Underlining Test subscales were associated with significant correlation coefficients. Consistency values for these subtests was .27 (Subtest 13) and less. A graphical display of the Pearson-r coefficients for these measures is provided in Figure 6. The nature of each subtest is provided in Appendix F. Rank values for these variables ranged from 33rd (Subtest 13) to 63rd (Subtest 9). The median ranking for the Underlining Test variables was found to be 54.

None of the Tactile-Perceptual measures yielded significant Pearson-r coefficients. Tactile (L) errors (r = .33) was ranked 26th, in terms of consistency, the highest for this ability domain. The remaining three measures were afforded ranks of 48th and lower. Median ranking for the entire domain was 49. Figure 7 graphically displays the reliability coefficients for these variables.

Finally, the sole measure of Right-Left Awareness was associated with a non-significant retest coefficient of -.13. The ranking of this measure was 61st, the lowest of all domains with respect to consistency.

In general, the neuropsychological measures

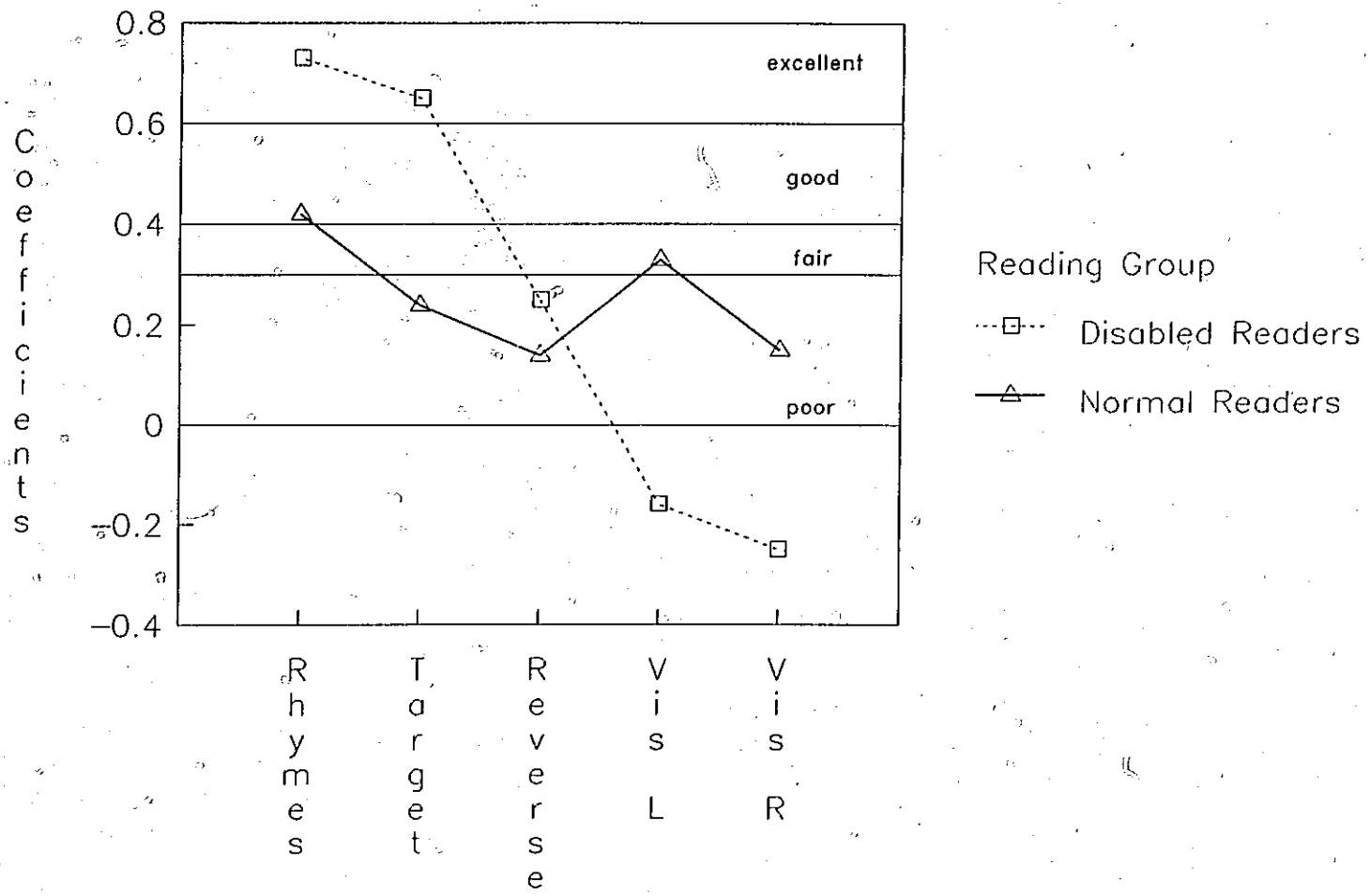

Figure 5. Consistency of visual–perceptual measures Disabled vs. Normal readers.

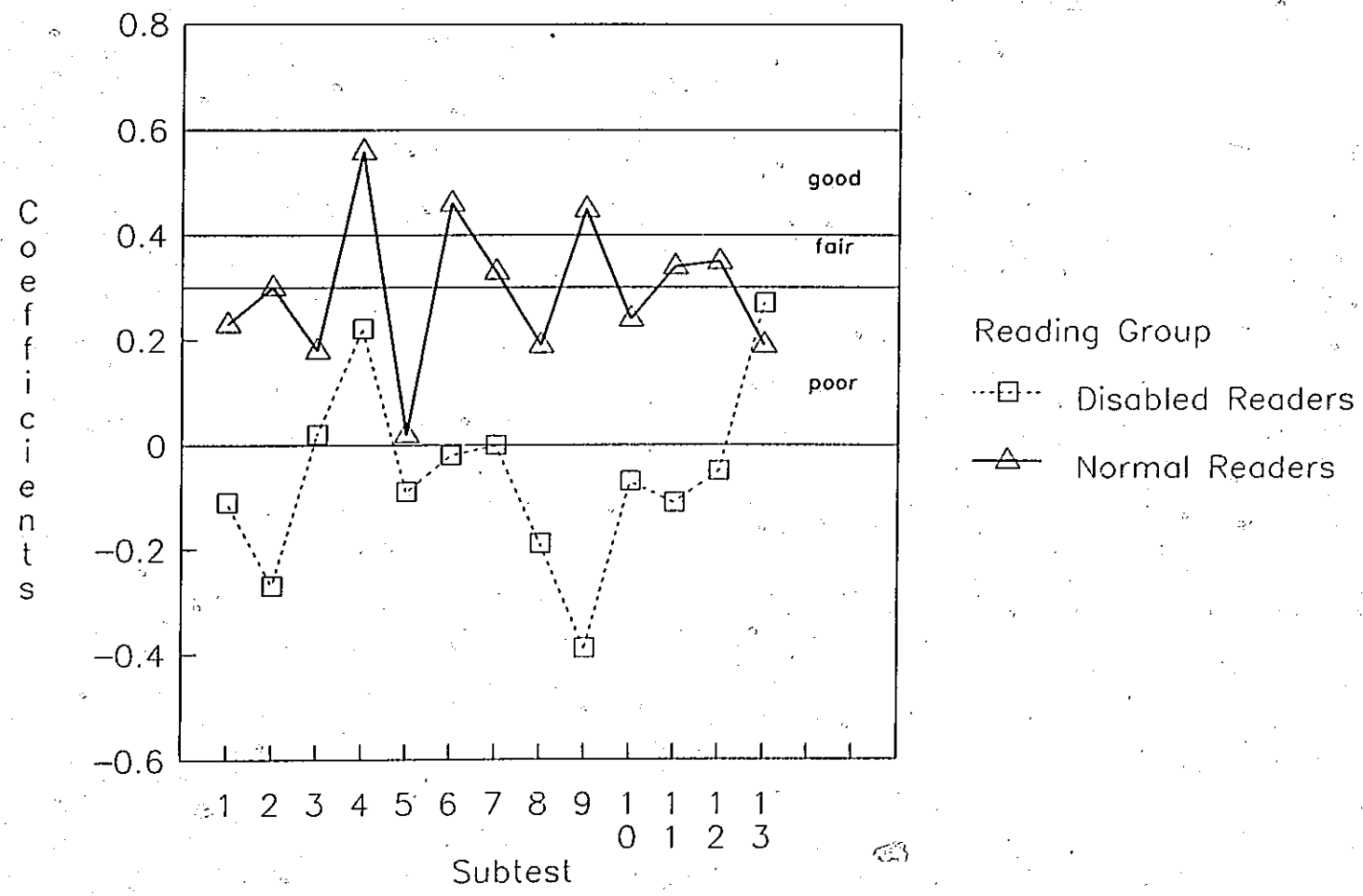

Figure 6. Consistency of Underlining test measures Disabled vs. Normal readers.

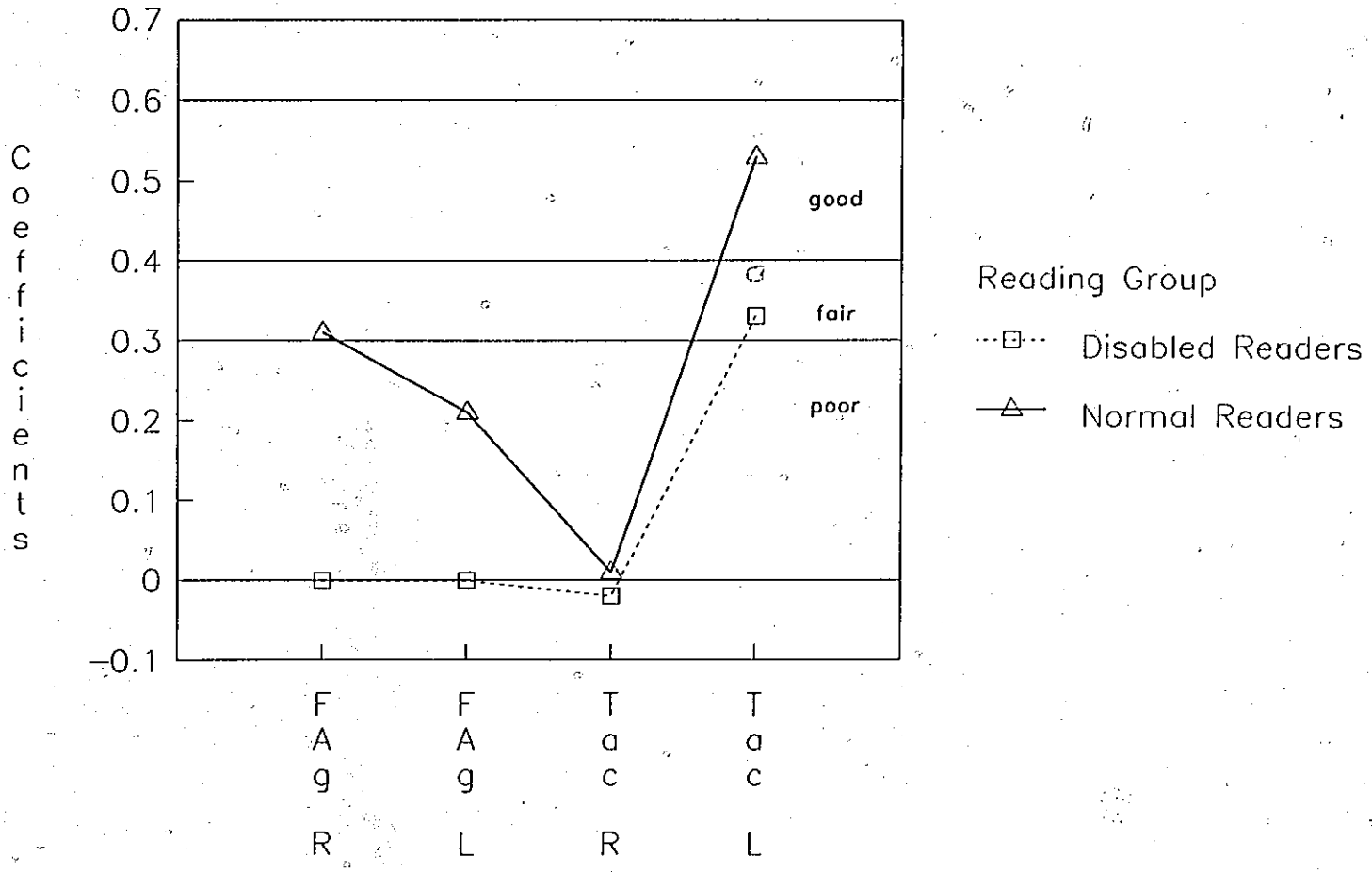

Figure 7. Consistency of tactile–perceptual measures. Disabled vs. Normal readers.

demonstrated a lesser degree of stability for the DR childrens' performance relative to their consistency. The ICC coefficients ranged in value from .47 (Target correct) to -.37 for Subtest 9 of the Underlining Test. The mean ICC value for all 63 neuropsychological measures was found to be .08 (SD = .18). Only six of the coefficients proved to be statistically significant to the $p < .05$ level. Twenty-three of the 63 measures were associated with coefficients of stability in the negative direction. Of these measures, nine reflected a general pattern of improvement in the childrens' performances over time: Name Writing speed (R), Name Writing (L) speed, Maze Time (L), Maze Counter (L) errors, Visual (L) errors, Visual (R) errors, Tactile (L) errors, Finger Agnosia (L) errors, and, Auditory (L) errors. Of the 41 measures assigned a positive ICC coefficient, six variables (Maze Speed (L), Maze Speed (R), Maze Counter (R) errors, Maze Time (R), Finger Agnosia (R) errors, and Tactile (R) errors, were indicative of a worsening degree of relationship between DR subjects' initial and subsequent test performance. Insufficient data precluded the calculation of an ICC coefficient for the Auditory (R) errors measure. Table 4 contains the ICC coefficients for these measures rank ordered in terms of their stability. The same coefficients, arranged for ability domain, separately, are presented in Table 5.

Table 4

Rank Order of Disabled Readers' Tests Based on Magnitude of Correlation (ICC)

Rank	Variable	ICC
1	Target correct	.47
2	Finger Tapping (R) taps/10"	.43
3	Maze Speed (L)	.41
4	Rhymes correct	.40
5	Similarities Scaled Score	.38
6	Finger Tapping (L) taps/10"	.38
7	Performance IQ	.33*
8	Verbal IQ	.27*
9	Vocabulary Scaled Score	.27*
10	Maze Counter (R) errors	.27*
11	Auditory Closure correct	.26*
12	Information Scaled Score	.26*
13	Underlining Subtest 13	.24*
14	Underlining Subtest 4	.22*
15	Underlining Subtest 11	.20*
16	Picture Arrangement Scaled Score	.22*
17	Block Design Scaled Score	.22*
18	Digit Span Scaled Score	.22*
19	Grip Strength (R) Kg.	.22*
20	MAT Word Knowledge Standard Score	.18*
21	Object Assembly Scaled Score	.17*
22	WRAT Reading Standard Score	.17*
23	Full Scale IQ	.17*
24	Reverse correct	.16*
25	Finger Agnosia (R) errors	.16*
26	WRAT Arithmetic Standard Score	.14*

Table 4 (Continued)

Rank	Variable	ICC
27	Grip Strength (L) Kg.	.13*
28	PPVT-IQ	.13*
29	Picture Completion Scaled Score	.12*
30	Maze Speed (R)	.11*
31	Sentence Memory correct	.10*
32	Verbal Fluency correct	.10*
33	WRAT Spelling Standard Score	.04*
34	Foot Tapping (R) taps/10"	.04*
35	Underlining Subtest 3	.02*
36	Foot Tapping (L) taps/10"	.02*
37	Maze Time (R)	.02*
38	Comprehension Scaled Score	.01*
39	Tactile (R) errors	.002*
40	Underlining Subtest 7	-.001*
41	Underlining Subtest 6	-.01*
42	Name Writing speed (R)	-.03*
43	Maze Counter (L) errors	-.04*
44	Auditory (L) errors	-.04*
45	Underlining Subtest 12	-.04*
46	Tactile (L) errors	-.05*
47	Speech-sounds Perception correct	-.05*
48	MAT Reading Standard Score	-.06*
49	Coding Scaled Score	-.06*
50	Underlining Subtest 10	-.06*
51	Visual (L) errors	-.06*
52	Finger Agnosia (L) errors	-.07*
53	MAT Word Discrimination Standard Score	-.08*
54	Right-Left Awareness correct	-.08*
55	Underlining Subtest 5	-.08*

Table 4 (Continued)

Rank	Variable	ICC
56	Underlining Subtest 1	-.10*
57	Underlining Subtest 8	-.13*
58	Visual (R) errors	-.16*
59	Arithmetic Scaled Score	-.20*
60	Name Writing speed (L)	-.20*
61	Maze Time (L)	-.21*
62	Underlining Subtest 2	-.25*
63	Underlining Subtest 9	-.37*
64	Auditory (R) errors	-- ++

* $p > .05$

++ Insufficient data to calculate ICC coefficient

Table 5

ICC's for Disabled Readers' Tests Within Ability Domains

Test Measure	ICC	
Psychometric Intelligence		
Similarities Scaled Score	.38	Fair
Performance IQ	.33	
Verbal IQ	.27	Poor
Vocabulary Scaled Score	.27	
Information Scaled Score	.26	
Picture Arrangement Scaled Score	.22	
Block Design Scaled Score	.22	
Digit Span Scaled Score	.22	
Object Assembly Scaled Score	.17	
Full Scale IQ	.17	
Picture Completion Scaled Score	.12	
Comprehension Scaled Score	.01	
Coding Scaled Score	-.06	
Arithmetic Scaled Score	-.20	
Academic Achievement and Reading		
MAT Word Knowledge Standard Score	.18	Poor
WRAT Reading Standard Score	.17	
WRAT Arithmetic Standard Score	.14	
WRAT Spelling Standard Score	.04	
MAT Reading Standard Score	-.06	
MAT Word Discrimination Standard Score	-.08	
Auditory-Perception and Language		
Auditory Closure correct	.26	Poor
PPVT-IQ	.13	
Sentence Memory correct	.10	
Verbal Fluency correct	.10	
Auditory (L) errors	-.04	

Table 5 (Continued)

Test Measure	ICC	
Speech-sounds Perception correct	-.05	Poor
Auditory (R) errors	-- ++	
Motor		
Finger Tapping (R) taps/10"	.43	Good
Maze Speed (L)	.41	
Finger Tapping (L) taps/10"	.38	Fair
Maze Counter (R) errors	.27	Poor
Grip Strength (R) Kg.	.22	
Grip Strength (L) Kg.	.13	
Maze Speed (R)	.11	
Foot Tapping (R) taps/10"	.04	
Foot Tapping (L) taps/10"	.02	
Maze Time (R)	.02	
Name Writing speed (R)	-.03	
Maze Counter (L) errors	-.04	
Name Writing speed (L)	-.20	
Maze Time (L)	-.21	
Visual-Perception		
Target correct	.47	Good
Rhymes correct	.40	
Reverse correct	.16	Poor
Visual (L) errors	-.06	
Visual (R) errors	-.16	
Tactile-Perception		
Finger Agnosia (R) errors	.16	Poor
Tactile (R) errors	.002	
Tactile (L) errors	-.05	
Finger Agnosia (L) errors	-.07	

Table 5 (Continued)

Test Measure	ICC	
Underlining Test		
Underlining Subtest 13	.24	Poor
Underlining Subtest 4	.22	
Underlining Subtest 11	.22	
Underlining Subtest 3	.02	
Underlining Subtest 7	-.001	
Underlining Subtest 6	-.01	
Underlining Subtest 12	-.04	
Underlining Subtest 10	-.06	
Underlining Subtest 5	-.08	
Underlining Subtest 1	-.10	
Underlining Subtest 8	-.13	
Underlining Subtest 2	-.25	
Underlining Subtest 9	-.37	
Right-Left Awareness		
Right-Left Awareness correct	-.08	Poor

++ Insufficient data to calculate ICC coefficient

Within the Psychometric Intelligence ability sphere, coefficient values ranged from .38 (Similarities subtest) to -.20 (Arithmetic subtest). A plot of these results is provided in Figure 8. Eleven of the 14 measures examined occupied ranks that fell within the upper half of the distribution; however, none was ranked greater than 5th for stability. The median rank was found to be 18 for these variables. In addition to the Arithmetic subtest measure, a negative ICC coefficient value was found for the childrens' performance on the Coding subtest (r = -.06).

The range of stability coefficient values for the Visual-Perceptual measures was from .47 (Target correct) to -.16 (Visual (R) errors). In terms of their rankings for stability, three of the five variables occupied positions within the upper third of the rank order. Target correct and Rhymes correct were ranked 1st and 4th, respectively. Reversals correct proved to be the median measure and was ranked 24th. A plot of the results for the Visual-Perceptual measures is provided in Figure 9.

The six measures of Academic Achievement and Reading ability ranged in terms of their ICC coefficient values from .18 (MAT Word Knowledge subtest) to -.08 (MAT Word Discrimination subtest). Figure 10 contains these results in a graphical format. These variables were roughly divided between the upper and lower halves of the rank order distribution (ranks ranging from 20th to 53rd). The median

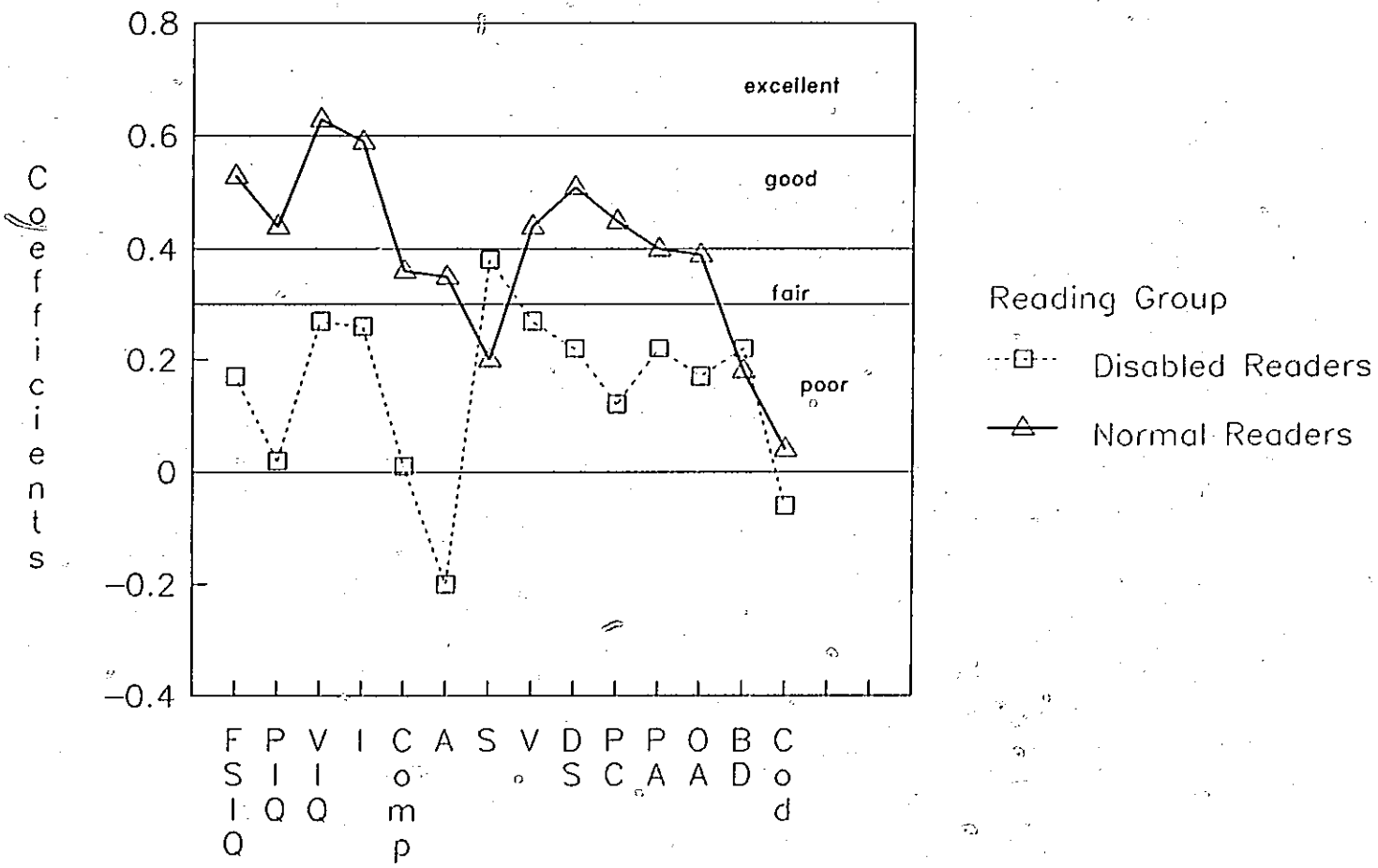

Figure 8. Stability of psychometric intelligence measures. Disabled vs. Normal readers.

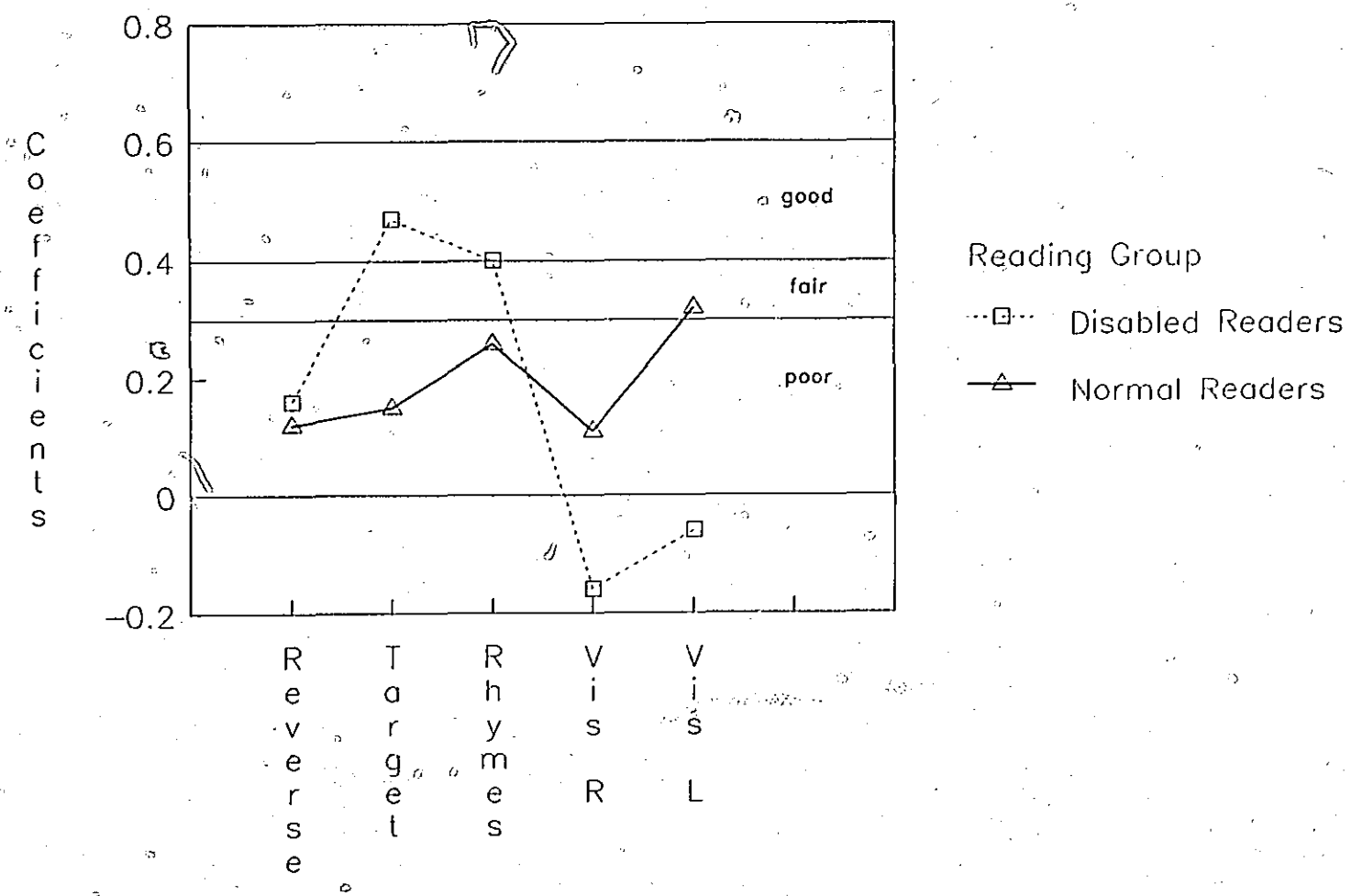

Figure 9. Stability of visual-perceptual measures. Disabled vs. Normal readers.

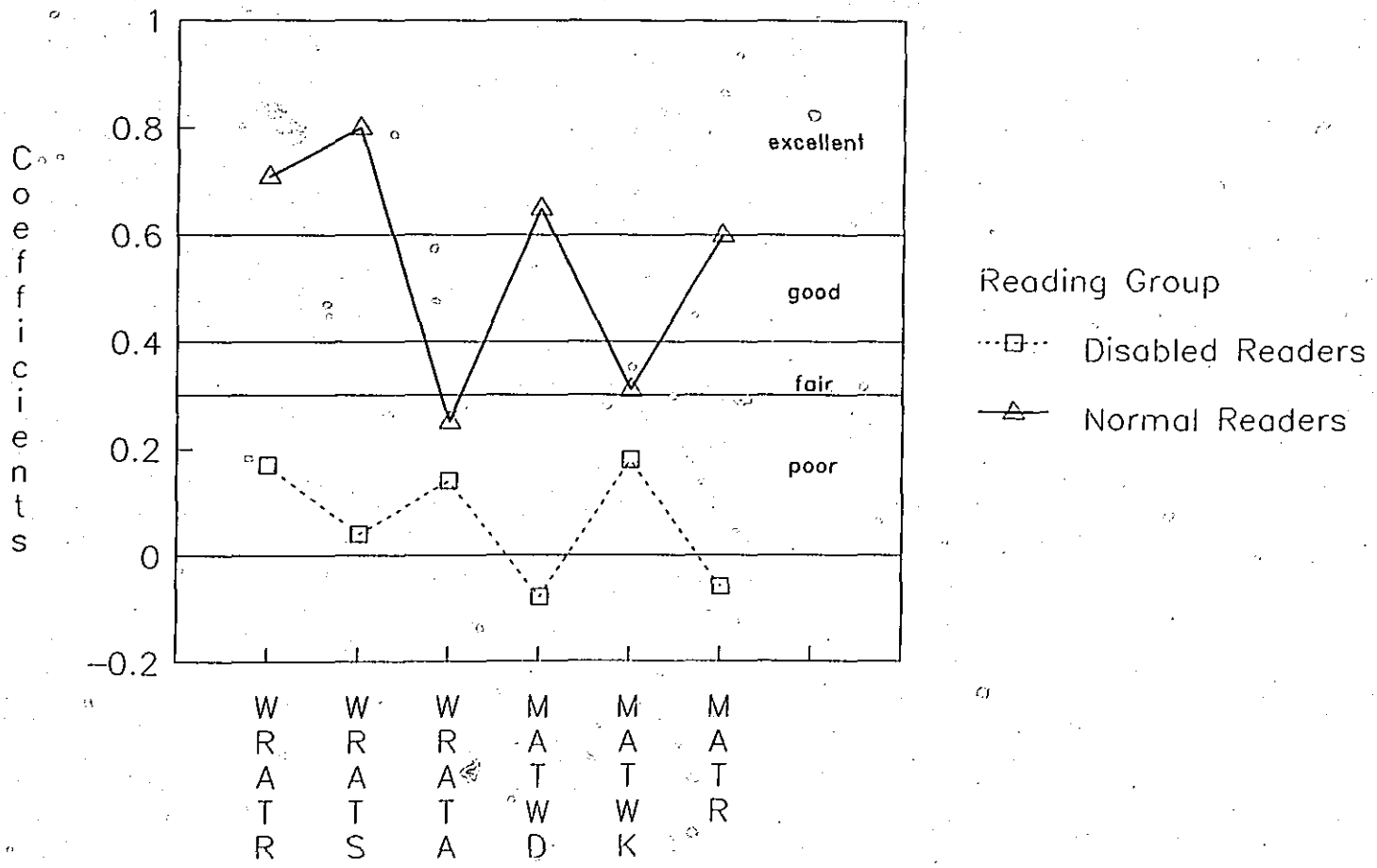

Figure 10. Stability of academic achievement and reading measures. Disabled vs. Normal readers.

ability rank found to be 30.

For the six measures of Auditory-Perception and Language ability, ICC coefficients ranged from -.26 for Auditory Closure correct, to -.05 for Speech-sounds Perception correct. With the exception of the former measure, which was ranked 11th in terms of its stability, the remaining measures occupied positions roughly within the middle of the rank order distribution. Rankings for these measures ranged from 28th for PPVT-IQ to 47th for Speech-sounds Perception correct. Median rank value for all Auditory-Perception and Language variables was 31. Figure 11 presents a graph of these ICC coefficients.

A considerable amount of variation was seen for the ICC coefficients derived from the DR group's performance on the Motor ability measures. Values ranged from .43 (Finger Tapping (R)) to -.21 (Maze Time (L)). Two measures, Finger Tapping (R) and Maze Speed (L), were ranked 2nd and 3rd (respectfully) for stability. The measures of Name Writing speed (L) and Maze Time (L) received respective rankings of 60 and 61. Coefficients for all Motor measures are contained within the graph presented as Figure 12. The median rank associated with the Motor measures was 32.

The four measures of Tactile-Perceptual ability were ranked in the lower two-thirds of the stability rank order distribution. The median rank for these measures was found to be 43. A plot of these results are provided in Figure

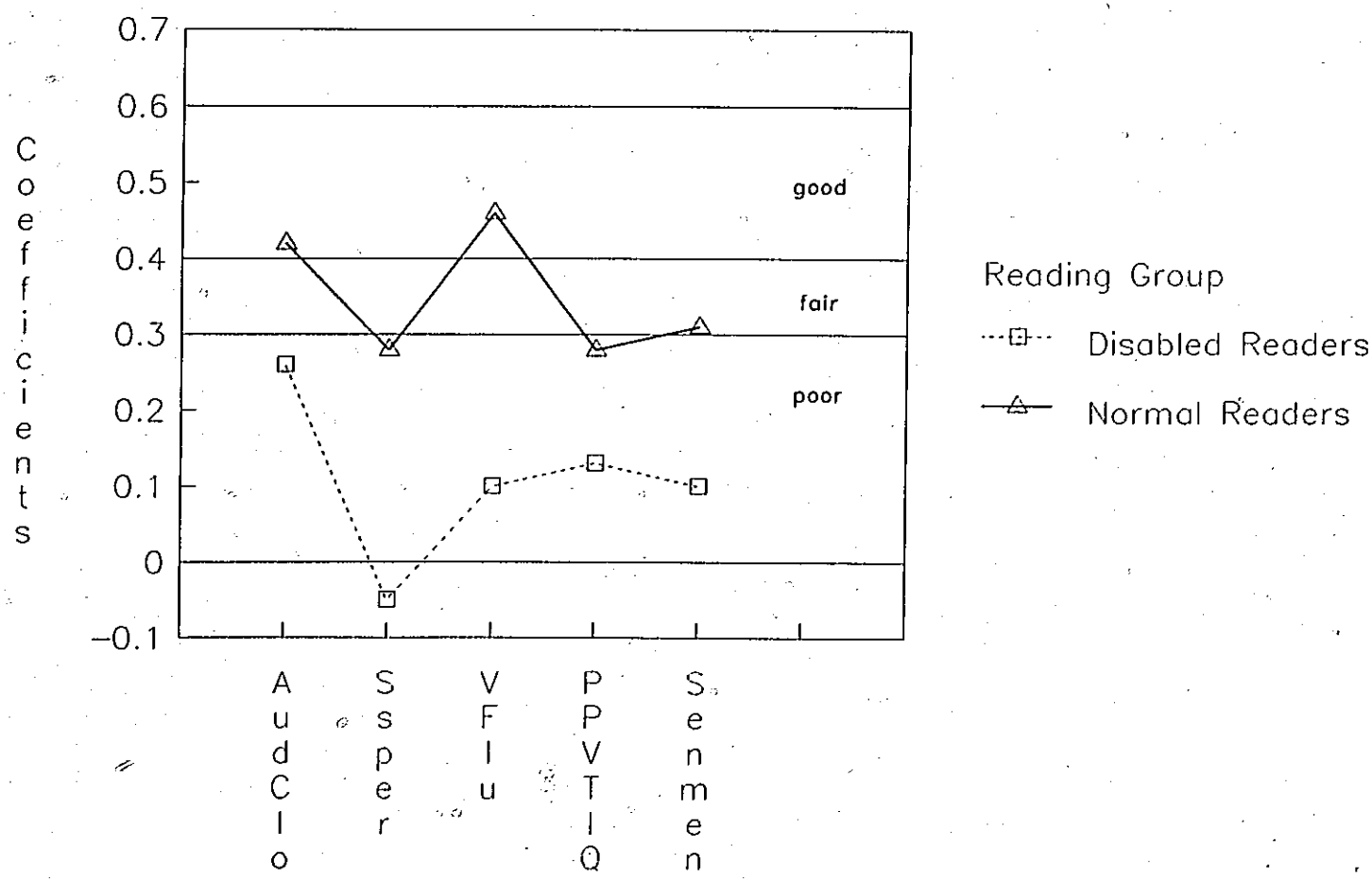

Figure 11. Stability of auditory–perceptual and language measures. Disabled vs. Normal readers.

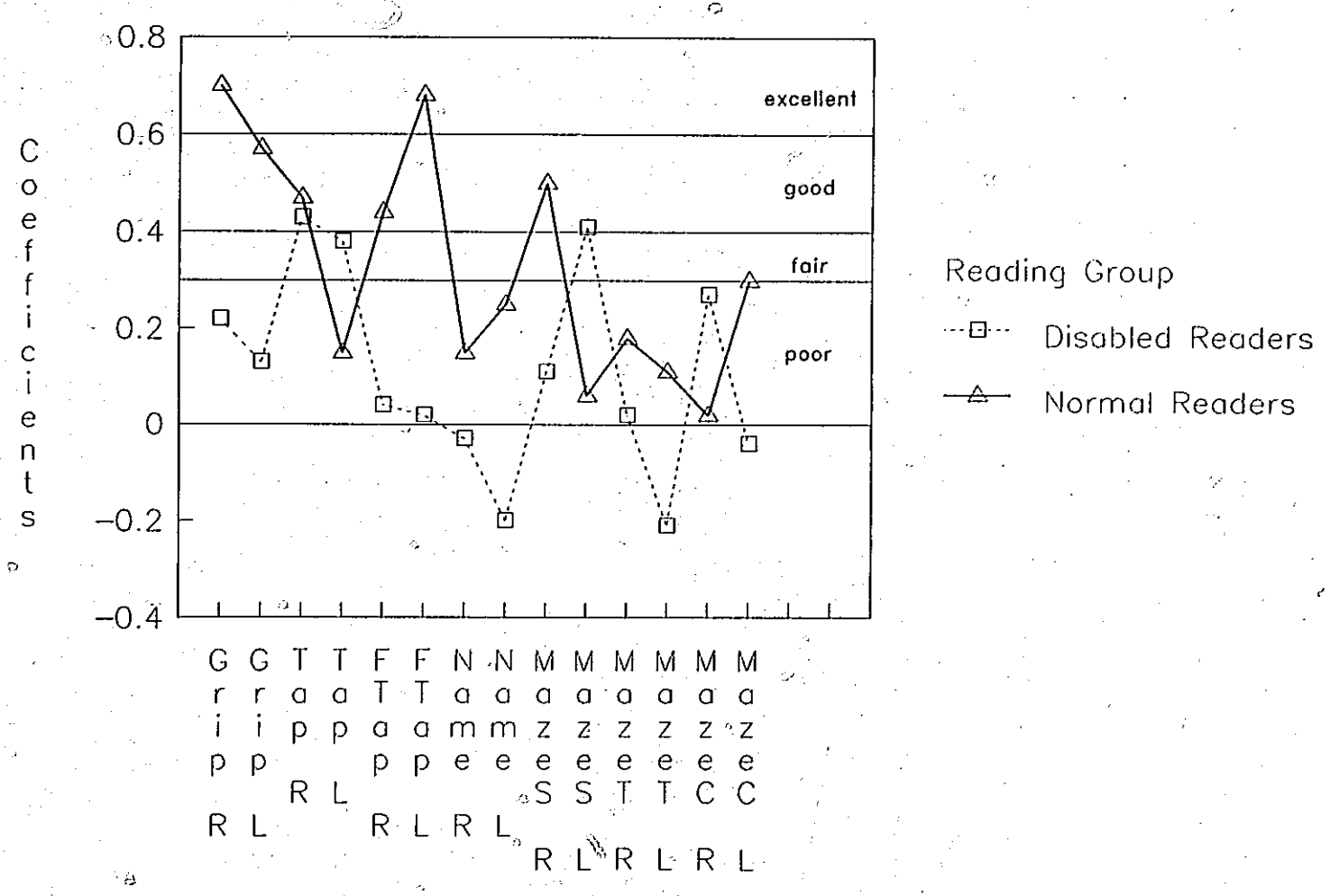

Figure 12. Stability of motor measures. Disabled vs. Normal readers.

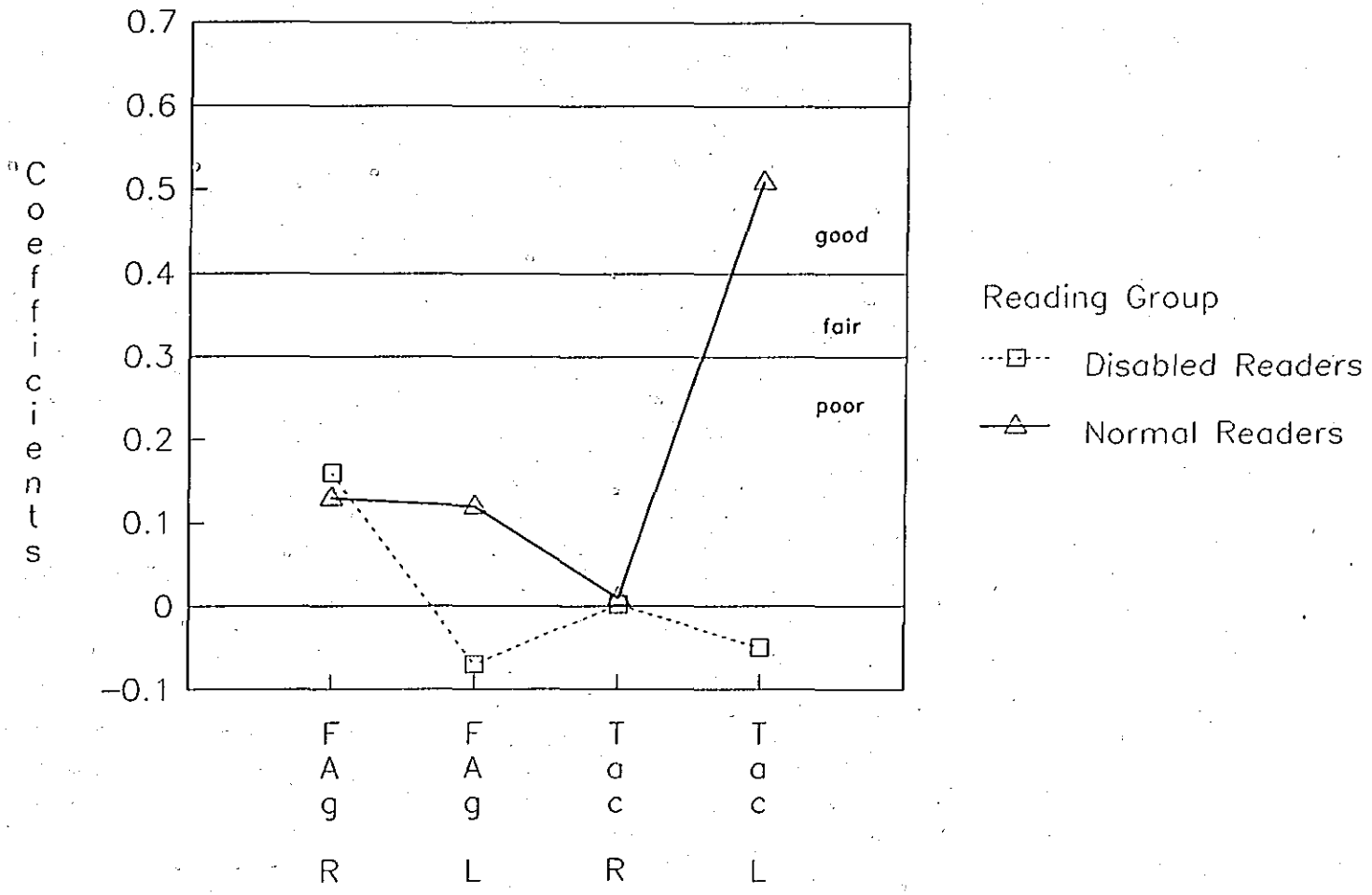

Figure 13. Stability of tactile-perceptual measures. Disabled vs. Normal readers.

13. Finger Agnosia (R) errors was associated with a coefficient of .16 and received the highest ranking of these measures, 25th. The least stable measure proved to be Finger Agnosia (L) errors ($\underline{r}$ = -.07), and this variable received a ranking of 52nd.

The Underlining Test subscales mostly occupied positions within the lower half of the rankings of the ICC coefficient values. The largest ICC values were evident for Subtest 13 ($\underline{r}$ = .24), Subtest 4 ($\underline{r}$ = .22), and Subtest 11 ($\underline{r}$ = .20). These measures had stability rankings of 13, 14, and 15, respectively. The remaining subscales ranged in ICC value from .02 (Subtest 3) to -.37 (Subtest 9), and were ranked from 35th to 63rd, respectively. Figure 14 presents a graphical representation of these results.

The single Right-Left Awareness measure was associated with an ICC coefficient value of -.08. The ranking found for this variable was 54th, in terms of its stability.

Investigation 2

An examination of the consistency and stability of the NR childrens' two-year retest performance on the 64 neuropsychological variables was undertaken in this investigation. The means and standard deviations for the NR childrens' performance on each of the variables is presented in Table 6. The resultant rank orders of the Pearson-$\underline{r}$ and ICC coefficients for the NR group were compared to those found for the DR sample in Investigation 1. A Spearman rank

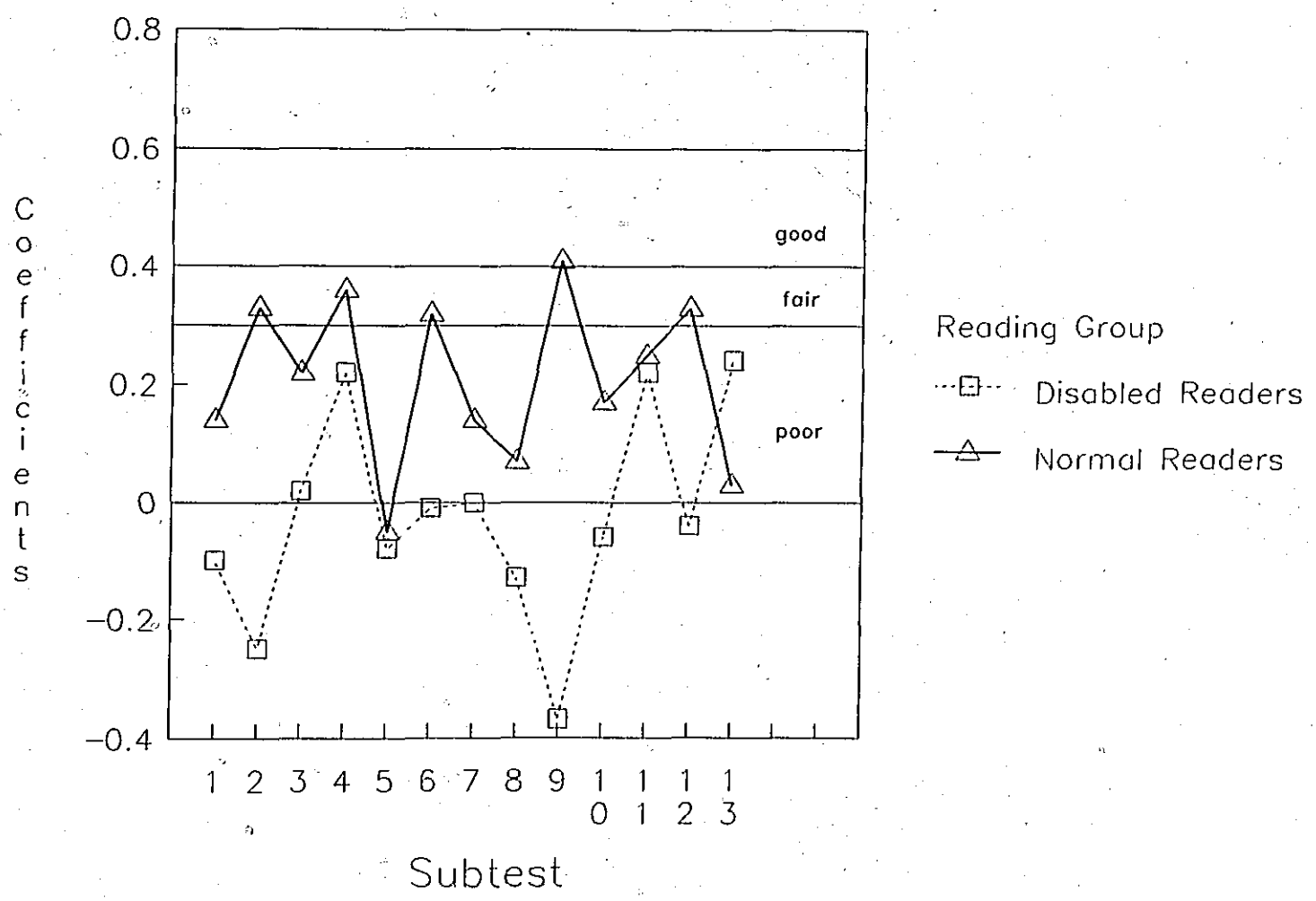

Figure 14. Stability of Underlining test measures. Disabled vs. Normal readers.

Table 6

Descriptive Statistics of Test Variables for Normal Readers

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Psychometric Intelligence				
Full Scale IQ	107.1	7.3	111.8	9.11
Verbal IQ	102.3	10.2	108.8	8.8
Performance IQ	111.1	8.6	112.9	11.5
Information Scaled Score	10.1	2.5	10.5	2.3
Comprehension Scaled Score	9.9	3.0	10.8	2.1
Digit Span Scaled Score	10.2	2.1	11.2	2.6
Arithmetic Scaled Score	11.6	2.6	12.1	2.5
Similarities Scaled Score	10.7	2.8	12.1	2.5
Vocabulary Scaled Score	9.5	2.4	11.6	2.4
Picture Completion Scaled Score	12.4	2.6	11.1	2.7
Picture Arrangement Scaled Score	11.0	2.6	13.0	2.1
Block Design Scaled Score	11.8	2.5	13.0	2.1
Object Assembly Scaled Score	11.7	2.3	12.1	2.9
Coding Scaled Score	11.0	2.1	11.8	2.5

Table 6 (Continued)

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Academic Achievement and Reading				
MAT Word Knowledge Standard Score	48.9	8.2	57.0	5.6
MAT Word Discrimination Standard Score	54.6	8.0	57.8	6.0
MAT Reading Standard Score	49.1	5.3	53.2	7.4
WRAT Reading Standard Score	117.7	15.2	122.3	17.9
WRAT Spelling Standard Score	112.3	14.4	112.8	17.9
WRAT Arithmetic Standard Score	100.3	4.4	102.7	8.0
Auditory-Perception and Language				
PPVT-IQ	109.0	11.7	111.5	10.6
Auditory Closure correct	13.2	3.8	15.7	3.7
Sentence Memory correct	12.5	2.5	14.4	2.4
Verbal Fluency correct	6.8	2.5	8.7	1.9
Auditory-Perception and Language				
Auditory (R) errors	.1	.2	-- *	-- *
Auditory (L) errors	.2	.9	.2	.6
Speech-sounds Perception correct	22.8	3.4	26.1	2.2

Table 6 (Continued)

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Motor				
Finger Tapping (R) taps/10"	30.1	4.7	34.0	5.1
Finger Tapping (L) taps/10"	28.1	3.1	31.1	4.2
Foot Tapping (R) taps/10"	25.6	4.9	32.0	4.6
Foot Tapping (L) taps/10"	24.2	4.0	30.0	4.3
Maze Speed (R)	113.5	34.1	107.1	34.4
Maze Speed (L)	109.1	27.8	97.9	21.6
Maze Time (R)	3.6	1.7	1.6	1.1
Maze Time (L)	8.1	3.8	5.4	2.7
Maze Counter (R) errors	26.6	10.8	13.1	7.3
Maze Counter (L) errors	53.3	19.7	39.6	14.5
Grip Strength (R) Kg.	12.6	2.7	13.9	3.2
Grip Strength (L) Kg.	11.9	2.2	12.5	2.8
Name Writing speed (R)	19.6	6.0	10.5	2.9
Name Writing speed (L)	32.8	10.8	24.3	7.4
Visual-Perception				
Target correct	14.8	2.8	17.7	2.0
Rhymes correct	16.7	3.1	18.9	1.4
Reverses correct	34.3	4.0	39.9	2.7

Table 6 (Continued)

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Visual (R) errors	.1	.2	.2	.6
Visual (L) errors	.1	.3	.1	.4
Tactile-Perception				
Finger Agnosia (R) errors	1.5	1.7	.9	1.2
Finger Agnosia (L) errors	1.6	1.8	.7	1.1
Tactile (R) errors	1.1	1.3	.3	.8
Tactile (L) errors	.6	1.1	.3	.9
Underlining Test				
Subtest 1	27.4	6.4	23.1	5.0
Subtest 2	31.2	6.8	25.6	4.0
Subtest 3	19.0	4.0	15.6	3.8
Subtest 4	8.5	5.7	7.8	4.4
Subtest 5	24.8	5.8	19.7	4.2
Subtest 6	16.9	4.6	15.1	2.7
Subtest 7	16.5	4.4	12.6	3.1
Subtest 8	9.9	4.0	7.5	1.3
Subtest 9	8.7	3.7	6.7	2.0
Subtest 10	15.8	6.4	13.6	3.8
Subtest 11	16.6	5.5	15.6	2.6

Table 6 (Continued)

	Year-0 Assessment		Year-2 Assessment	
	Mean	Standard Deviation	Mean	Standard Deviation
Subtest 12	8.9	2.5	7.5	1.8
Subtest 13	24.4	6.6	21.6	2.8
Right-Left Awareness				
Right-Left Awareness correct	20.0	5.1	22.1	4.3

* Insufficient data available

order correlation coefficient served as the measure of comparison.

Thirty-four of the Pearson-r coefficients yielded by the NR group's performance were statistically significant to the $p < .05$ level. These correlation coefficients ranged in value from .88 (Auditory (L) errors) to .38 (PPVT-IQ). The mean Pearson-r coefficient value was found to be .39 (SD = .21). All of the coefficients are provided in ranked order in Table 7. These same coefficient values are presented within each of the ability areas with which they are associated in Table 8.

Five of six measures of Academic Achievement and Reading yielded significant correlation coefficients. Two measures, the WRAT Spelling subtest ($r = .81$) and the WRAT Reading subtest ($r = .77$) were ranked 2nd and 3rd(respectively) on the basis of consistency. With the exception of the MAT Word Knowledge subtest ($r = .41$) which was ranked 49th in consistency, the remainder of the measures were ranked within the upper half of the rank order distribution. A plot of these results is provided in Figure 3. A rank value of 9 was found to be the median value for this ability domain.

Of the 14 measures sensitive to Psychometric Intelligence, 11 were associated with significant correlation coefficients. The WISC Full Scale IQ ($r = .67$) measure proved to be the most consistent variable. The

Table 7

Rank Order of Normal Readers' Tests Based on Magnitude of Correlation (Pearson-r)

Rank	Variable	Pearson-r
1	Auditory (L) errors	.88
2	WRAT Spelling Standard Score	.81
3	WRAT Reading Standard Score	.77
4	Grip Strength (R) Kg.	.74
5	Auditory Closure correct	.69
6	Foot Tapping (L) taps/10"	.69
7	Full Scale IQ	.67
8	MAT Reading Standard Score	.67
9	Verbal IQ	.66
10	MAT Word Discrimination Standard Score	.64
11	Grip Strength (L) Kg.	.63
12	Information Scaled Score	.57
13	Underlining Subtest 4	.56
14	Similarities Scaled Score	.53
15	Tactile (L) errors	.53
16	Digit Span Scaled Score	.53
17	Speech-sounds Perception correct	.52
18	Vocabulary Scaled Score	.50
19	Performance IQ	.49
20	Foot Tapping (R) taps/10"	.48
21	Maze Speed (R)	.48
22	Underlining Subtest 6	.46
23	Picture Arrangement Scaled Score	.45
24	Underlining Subtest 9	.45
25	Maze Counter (L) errors	.43
26	Rhymes correct	.42

Table 7 (Continued)

Rank	Variable	Pearson-r
27	MAT Word Knowledge Standard Score	.41
28	Name Writing speed (L)	.41
29	Arithmetic Scaled Score	.40
30	Verbal Fluency correct	.40
31	Comprehension Scaled Score	.40
32	Object Assembly Scaled Score	.40
33	Picture Comprehension Scaled Score	.39
34	PPVT-IQ	.38
35	Underlining Subtest 12	.35*
36	Underlining Subtest 11	.34*
37	Visual (L) errors	.33*
38	Underlining Subtest 7	.33*
39	Block Design Scaled Score	.32*
40	Finger Agnosia (R) errors	.31*
41	Sentence Memory correct	.31*
42	Underlining Subtest 2	.30*
43	Finger Tapping (R) taps/10"	.30*
44	Target correct	.24*
45	Underlining Subtest 10	.24*
46	Maze Time (L)	.23*
47	Underlining Subtest 1	.23*
48	Finger Agnosia (L) errors	.21*
49	WRAT Arithmetic Standard Score	.21*
50	Underlining Subtest 8	.19*
51	Name Writing speed (R)	.19*
52	Underlining Subtest 13	.19*
53	Maze Time (R)	.18*
54	Underlining Subtest 3	.18*
55	Right-Left Awareness correct	.17*

Table 7 (Continued)

Rank	Variable	Pearson-r
56	Visual (R) errors	.15*
57	Reverse correct	.14*
58	Finger Tapping (L) taps/10"	.12*
59	Maze Speed (L)	.09*
60	Coding Scaled Score	.09*
61	Maze Counter (R) errors	.02*
62	Underlining Subtest 5	.02*
63	Tactile (R) errors	.01*
64	Auditory (R) errors	-- ++

* $p > .05$

++ Insufficient data to calculate Pearson-r coefficient

Table 8

Pearson-r's for Normal Readers' Tests Within Ability Domains

Test Measure	Pearson-r	
Psychometric Intelligence		
Full Scale IQ	.67	Excellent
Verbal IQ	.66	
Information Scaled Score	.57	Good
Similarities Scaled Score	.53	
Digit Span Scaled Score	.53	
Vocabulary Scaled Score	.50	
Performance IQ	.49	
Picture Arrangement Scaled Score	.45	
Arithmetic Scaled Score	.40	
Comprehension Scaled Score	.40	
Object Assembly Scaled Score	.40	
Picture Comprehension Scaled Score	.39	Fair
Block Design Scaled Score	.32	
Coding Scaled Score	.09	Poor
Academic Achievement and Reading		
WRAT Spelling Standard Score	.81	Excellent
WRAT Reading Standard Score	.77	
MAT Reading Standard Score	.67	
MAT Word Discrimination Standard Score	.64	
MAT Word Knowledge Standard Score	.41	Good
WRAT Arithmetic Standard Score	.21	Poor
Auditory-Perception and Language		
Auditory (L) errors	.88	Excellent
Auditory Closure correct	.69	
Speech-sounds Perception correct	.52	Good
Verbal Fluency correct	.40	

Table 8 (Continued)

Test Measure	Pearson-r	
PPVT-IQ	.38	Fair
Sentence Memory correct	.31	
Auditory (R) errors	-- ++	
Motor		
Grip Strength (R) Kg.	.74	Excellent
Foot Tapping (L) taps/10"	.69	
Grip Strength (L) Kg.	.63	
Foot Tapping (R) taps/10"	.48	Good
Maze Speed (R)	.48	
Maze Counter (L) errors	.43	
Name Writing speed (L)	.41	
Finger Tapping (R) taps/10"	.30	Fair
Maze Time (L)	.23	Poor
Name Writing speed (R)	.19	
Maze Time (R)	.18	
Finger Tapping (L) taps/10"	.12	
Maze Speed (L)	.09	
Maze Counter (R) errors	.02	
Visual-Perception		
Rhymes correct	.42	Good
Visual (L) errors	.33	Fair
Target correct	.24	Poor
Visual (R) errors	.15	
Reverse correct	.14	
Tactile-Perception		
Tactile (L) errors	.53	Good

Table 8 (Continued)

Test Measure	Pearson-r	
Finger Agnosia (R) errors	.31	Fair
Finger Agnosia (L) errors	.21	Poor
Tactile (R) errors	.01	
Underlining Test		
Underlining Subtest 4	.56	Good
Underlining Subtest 6	.46	
Underlining Subtest 9	.45	
Underlining Subtest 12	.35	Fair
Underlining Subtest 11	.34	
Underlining Subtest 7	.33	
Underlining Subtest 2	.30	
Underlining Subtest 10	.24	Poor
Underlining Subtest 1	.23	
Underlining Subtest 8	.19	
Underlining Subtest 13	.19	
Underlining Subtest 3	.18	
Underlining Subtest 5	.02	
Right-Left Awareness		
Right-Left Awareness correct	.17	Poor

++ Insufficient data to calculate Pearson-r coefficient

least consistency was observed for the Picture Completion subtest measure (r = .39). A plot of the coefficients for the Psychometric Intelligence measures is presented in Figure 4. The majority of the Psychometric Intelligence variables received rankings in the upper half of the rank order distribution, and the median consistency ranking for this ability area was 20.

For those measures of Auditory-Perception and Language that yielded significant Pearson-r coefficients, the range of consistency was from .88 (Auditory (L) errors) to .38 (PPVT-IQ). None of the Auditory-Perception and Language measures received a ranking lower than 41; and, the median rank for these measures was observed to be 24. A plot ofthese results is provided in Figure 3.

Seven of the 14 Motor ability measures were associated with significant Pearson-r coefficients, ranging from .74 (Grip Strength (R)) to .41 (Name Writing speed (L)). The consistency rankings of these variables ranged from 4th for the former measure, to 61st for Maze Counter (R) errors. Figure 1 presents a plot of the Pearson-r coefficients for these variables. Median group consistency rank was found to be 36.

Fewer than one-quarter of the Underlining Test subtests were associated with significant consistency coefficients. The range in magnitude of these coefficients was from .57 (Subtest 4) to .45 (Subtest 9). None of the rankings for

these measures exceeded 13, with the vast majority of the variables being rated well within the lower half of the rank order distribution. The median Underlining Test consistency rank was found to be 42. Figure 6 presents the pattern of coefficients for these measures.

Finally, only one measure each, of the Visual-Perceptual and Tactile-Perceptual ability areas generated significant Pearson-*r* correlation coefficients. In the former case the Rhymes correct measure ($r = .42$) was ranked 26th in terms of consistency. The Tactile-Perceptual variable of Tactile (L) errors was associated with a coefficient of .53, and received a ranking of 15th. The single measure of Right-Left Awareness failed to produce a significant correlation coefficient, and was given a ranking of 55. A plot of the Visual-Perceptual and Tactile-Perceptual results are presented in Figures 5 and 7, respectively.

Overall, the stability (inferred from the ICC coefficient magnitudes) of the NR group's retest performance on the neuropsychological measures was quite variable. Thirty-three of the 64 measures were associated with significant ($p < .05$) ICC coefficients. The greatest magnitude of an ICC coefficient was .80, found for the Auditory (L) errors measure. Subtest 5 of the Underlining Test was observed to have the least stability. The ICC coefficient associated with the latter measure was

-.05, and was the only negative coefficient. The mean ICC coefficient for the NR childrens' performance was .33 (sd = .21). Table 9 contains the ICC coefficients calculated for all measures, ranked in terms of magnitude. The rankings of coefficients within ability areas is provided in Table 10.

The childrens' retest performance on 14 of the variables (Visual (R) errors, Visual (L) errors, Auditory (L) errors, Tactile (R) errors, Tactile (L) errors, Finger Agnosia (R) errors, Finger Agnosia (L) errors, Name Writing speed (R), Name Writing speed (L), Maze Counter (R) errors, Maze Counter (L) errors, Maze Time (R), Maze Time (L), Maze Speed (R), Maze Speed (L)) was indicative of a general decrement in performance for this group.

Four of the six measures of Academic Achievement and Reading were found to be ranked within the top 8 for stability. The ICC coefficient associated with the WRAT Spelling subtest equalled .80, while that related to the WRAT Arithmetic subtest earned the lowest coefficient value, .25. A plot of this data is presented in Figure 10. These variables ranged in terms of their rankings from 2 to 40, respectively. A median rank of 7 was found for all six Academic Achievement and Reading measures.

The ICC coefficients associated with the Psychometric Intelligence measures varied from a high value of .63 (WISC Verbal IQ) to .04 (Coding subtest). Figure 8 contains a graphical display of these coefficients. The median ranking

Table 9

Rank Order of Normal Readers' Tests Based on Magnitude of Correlation (ICC)

Rank	Variable	ICC
1	Auditory (L) errors	.80
2	WRAT Spelling Standard Score	.80
3	WRAT Reading Standard Score	.71
4	Grip Strength (R) Kg.	.70
5	Foot Tapping (L) taps/10"	.68
6	MAT Word Discrimination Standard Score	.65
7	Verbal IQ	.63
8	MAT Reading Standard Score	.60
9	Information Scaled Score	.59
10	Grip Strength (L) Kg.	.57
11	Full Scale IQ	.53
12	Digit Span Scaled Score	.51
13	Tactile (L) errors	.51
14	Maze Speed (R)	.50
15	Finger Tapping (R) taps/10"	.47
16	Verbal Fluency correct	.46
17	Picture Completion Scaled Score	.45
18	Foot Tapping (R) taps/10"	.44
19	Vocabulary Scaled Score	.44
20	Performance IQ	.44
21	Auditory Closure correct	.42
22	Underlining Subtest 9	.41
23	Picture Arrangement Scaled Score	.40
24	Object Assembly Scaled Score	.39
25	Underlining Subtest 4	.36

Table 9 (Continued)

Rank	Variable	ICC
26	Comprehension Scaled Score	.36
27	Arithmetic Scaled Score	.35
28	Underlining Subtest 2	.33
29	Underlining Subtest 12	.33
30	Underlining Subtest 6	.32
31	Visual (L) errors	.32
32	Sentence Memory correct	.31
33	MAT Word Knowledge Standard Score	.31
34	Maze Counter (L) errors	.30*
35	Speech-sounds Perception correct	.28*
36	PPVT-IQ	.28*
37	Rhymes correct	.26*
38	Underlining Subtest 11	.25*
39	Name Writing speed (L)	.25*
40	WRAT Arithmetic Standard Score	.25*
41	Underlining Subtest 3	.22*
42	Similarities Scaled Score	.20*
43	Block Design Scaled Score	.18*
44	Maze Time (R)	.18*
45	Underlining Subtest 10	.17*
46	Right-Left Awareness correct	.16*
47	Name Writing speed (R)	.15*
48	Target correct	.15*
49	Finger Tapping (L) taps/10"	.15*
50	Underlining Subtest 1	.14*
51	Underlining Subtest 7	.14*
52	Finger Agnosia (R) errors	.13*
53	Reverse correct	.12*
54	Finger Agnosia (L) errors	.12*
55	Maze Time (L)	.11*

Table 9 (Continued)

Rank	Variable	ICC
56	Visual (R) errors	.11*
57	Underlining Subtest 8	.07*
58	Maze Speed (L)	.07*
59	Coding Scaled Score	.04*
60	Underlining Subtest 13	.03*
61	Maze Counter (R) errors	.02*
62	Tactile (R) errors	.01*
63	Underlining Subtest 5	-.05*
64	Auditory (R) errors	-- ++

* $p > .05$

++ Insufficient data to calculate ICC coefficient

Table 10

ICC's for Normal Readers' Tests Within Ability Domains

Test Measure	ICC	
Psychometric Intelligence		
Verbal IQ	.63	Excellent
Information Scaled Score	.59	Good
Full Scale IQ	.53	
Digit Span Scaled Score	.51	
Picture Completion Scaled Score	.45	
Vocabulary Scaled Score	.44	
Performance IQ	.44	
Picture Arrangement Scaled Score	.40	
Object Assembly Scaled Score	.39	Fair
Comprehension Scaled Score	.36	
Arithmetic Scaled Score	.35	
Similarities Scaled Score	.20	Poor
Block Design Scaled Score	.18	
Coding Scaled Score	.04	
Academic Achievement and Reading		
WRAT Spelling Standard Score	.80	Excellent
WRAT Reading Standard Score	.71	
MAT Word Discrimination Standard Score	.65	
MAT Reading Standard Score	.60	
MAT Word Knowledge Standard Score	.31	Fair
WRAT Arithmetic Standard Score	.25	Poor
Auditory-Perception and Language		
Auditory (L) errors	.80	Excellent
Verbal Fluency correct	.46	Good
Auditory Closure correct	.42	

Table 10 (Continued)

Test Measure	ICC	
Sentence Memory correct	.31	Fair
Speech-sounds Perception correct	.28	Poor
PPVT-IQ	.28	
Auditory (R) errors	-- ++	
<u>Motor</u>		
Grip Strength (R) Kg.	.70	Excellent
Foot Tapping (L) taps/10"	.68	
Grip Strength (L) Kg.	.57	Good
Maze Speed (R)	.50	
Finger Tapping (R) taps/10"	.47	
Foot Tapping (R) taps/10"	.44	
Maze Counter (L) errors	.30	Fair
Name Writing speed (L)	.25	Poor
Maze Time (R)	.18	
Name Writing speed (R)	.15	
Finger Tapping (L) taps/10"	.15	
Maze Time (L)	.11	
Maze Speed (L)	.07	
Maze Counter (R) errors	.02	
<u>Visual-Perception</u>		
Visual (L) errors	.32	Fair
Rhymes correct	.26	Poor
Target correct	.15	
Reverse correct	.12	
Visual (R) errors	.11	
<u>Tactile-Perception</u>		
Tactile (L) errors	.51	Good
Finger Agnosia (R) errors	.13	Poor

Table 10 (Continued)

Test Measure	ICC	
Finger Agnosia (L) errors	.12	Poor
Tactile (R) errors	.01	
<u>Underlining Test</u>		
Underlining Subtest 9	.41	Good
Underlining Subtest 4	.36	Fair
Underlining Subtest 2	.33	
Underlining Subtest 12	.33	
Underlining Subtest 6	.32	
Underlining Subtest 11	.25	Poor
Underlining Subtest 3	.22	
Underlining Subtest 10	.17	
Underlining Subtest 1	.14	
Underlining Subtest 7	.14	
Underlining Subtest 8	.07	
Underlining Subtest 13	.03	
Underlining Subtest 5	-.05	
<u>Right-Left Awareness</u>		
Right-Left Awareness correct	.16	Poor

++ Insufficient data to calculate ICC coefficient

for these 14 measures was 22, reflecting the concentration of rankings of the variables' ranks within the upper half of the rank order distribution. Rank values ranged from 7 for Verbal IQ to 59 for the Coding subtest.

Regarding the stability of the childrens' performance over time, the Auditory-Perception and Language measures varied in terms of their ICC coefficients, from .80 (Auditory (L) errors) to .28 (PPVT-IQ). A coefficient for the Auditory (R) errors measure could not be calculated due to insufficient data. The median rank value for these measuring was 26. Figure 11 contains a plot of the ICC coefficients associated with the Auditory-Perception and Language measures.

Examining the measures within the Motor ability area, a great deal of variance was also observed for the ICC coefficients. The greatest stability was for Grip Strength (R), which had an ICC coefficient of .74. The least degree of stability was in evidence for the Maze Counter (R) errors measures ($\underline{r}$ = .02). Figure 12 presents a graphical display of the coefficients for all 14 Motor variables. The median stability rank value for this domain was 36.

The 13 measures that make up the Underlining Test ranged in ICC coefficient value from .41 (Subtest 9) to -.05 for (Subtest 5), and are presented in a graphical format in Figure 12. None of the Underlining subtest measures were ranked within the top one-third of the rank order

distribution, in terms of their stability. The median rank value for these measures was found to be 41.

Only one measure of Tactile-Perceptual was ranked within the upper half of the stability rank order. Tactile (L) errors ($\underline{r}$ = .51) received a rank of 12 regarding stability. In contrast however was the range of the remaining three measures within this ability domain. Values ranged from .13 (Finger Agnosia (R) errors) to .01 (Tactile (R) errors). The general concentration of these variables was within the lower one-third of the rank order distribution. A median rank value of 53 was observed within this domain. Figure 13 presents a plot of the coefficients associated with the Tactile-Perceptual variables.

Similar to the Tactile-Perceptual measures, all four of the Visual-Perceptual measures occupied ranks within the lower half of the rank order distribution. A rank of 31 was afforded to the Visual (L) errors ($\underline{r}$ = .32) variable. The least stable measure was (Visual (R) errors which was associated with a coefficient of value .11 (rank equals 56). The median rank value for these measures was 48.

Finally, the Right-Left Awareness measure was found to have a ICC coefficient of .16, and was ranked 46th in regards to its stability.

To compare the resulting rank orders of the NR group's Pearson-$\underline{r}$ and ICC coefficients with those for the DR children's performance, two Spearman rank order correlation

coefficients were calculated. To accomplish this task, the NONPARR CORR program in the SPSSx (Nie, 1983) batch system was employed. Comparing the rank orders of the Pearson-r coefficients between the two groups, the comparison was non-significant ($r(63) = .24$, $p > .05$). The second rank order correlation, comparing both groups' ordering of their respective ICC coefficients, also proved to be non-significant ($r(63) = .09$, $p > .10$).

Investigation 3

This third investigation compared the pattern of retest stability and consistency of the DR childrens' performance with that of two heterogeneous clinical samples on the series of neuropsychological measures. The rank order of the Pearson-r coefficients derived from the DR sample performance was compared to a similar rank order drawn from the Donaghy (1988) study. Table 11 contains the demographic characteristics of the heterogeneous clinical sample used in the Donaghy (1988) study. Only those neuropsychological measures that were included and administered in a similar fashion across the two studies were utilized for this analysis. As a result, the following measures that constitute a part of the current study were not included in subsequent analysis: The MAT Word Knowledge, Word Discrimination, and Reading subtests, Rhymes correct, Reverses correct, Right-Left Awareness correct, and the Underlining Test (Subtests 1 through 13). A Spearman rank

order correlation coefficient, calculated to compare the two rank orders, was highly significant ($r(45) = .70$, $p < .001$).

A similar analysis was next carried out to compare the DR childrens' rank ordering of the ICC coefficients, with a similar rank order drawn from the Brown (1987) study. The demographics characterizing the clinical sample from the Brown (1987) study are presented in Table 11. As in the previous comparison, only those tests that were common to both studies were included in the analysis. This condition necessitated the restriction of the same neuropsychological measures mentioned earlier in the context of the Donaghy (1988) comparison from inclusion in the current analysis. The result of the Spearman rank order correlation coefficient comparing the Brown (1987) and DR rank orders revealed that a strong relationship existed between the two orders' ICC rankings, $r(45) = .48$, $p < .001$.

Investigation 4

In this, the final investigation, the two-year retest consistency and stability of a pooled sample of DR and NR subjects was examined twice, over two separate two-year intervals. Following these examinations, the resulting rank orders were compared to one another for both, the Pearson-r and ICC coefficients, so as to provide a partial validation of the stability and consistency of these childrens' retest performance. The neuropsychological measures utilized in this investigation are included in Appendix F.

Table 11

Characteristics of Subject Samples used in Brown (1987) and Donaghy (1988) Studies

Characteristic	Brown (1987)	Donaghy (1988)
Number of Subjects:	248	322
Mean Age:	10.6 years	10.6 years
(SD):	(2.3)	(2.3)
Sex: Male	201	256
Female	45	66
Unclassified	2	0
Diagnostic Classification*		
Learning Disabled	131	167
Mentally Retarded	87	119
Brain Lesioned	30	55
Emotionally Disturbed	27	64
Environmentally Deprived	3	3
Retest Interval:	2.65 years	2.62 years
(SD):	(1.74)	(1.73)

* Some subjects were multiply classified

The first retest comparison occurred over the initial two years of the original longitudinal study. The children were assessed at the onset of the study (Year-0), and a further two-year period later (Year-2). The additional retest comparison interval was conducted between the childrens' second assessment (Year-2) and again two-years later still (Year-4). The means and standard deviations for the pooled sample's performance at the Year-0, Year-2, and Year-4 assessments are provided in Table 12.

The presentation of the results of this fourth investigation will be as follows. The consistency and stability observed in the performances of the pooled sample during the initial two-year retest interval will be presented first. This presentation will then be followed by a similar one examining the consistency and stability for these same children during the subsequent two-year retest interval. Finally, the results of two Spearman rank order correlation coefficients that directly compared the pattern of reliability and consistency between the first and second retest intervals will be offered.

Year-0/Year-2 Two-Year Retest Interval

Fifty-one of the 63 Pearson-*r* coefficients calculated on the pooled sample's initial two-year retest performance on the neuropsychological measures proved to be of a statistically significant level ($p < .05$). The range of the magnitude of the significant Pearson-*r* coefficients was from

Table 12

Descriptive Statistics of Test Variables for Pooled Sample

	Year-0 Assessment Mean (SD)	Year-2 Assessment Mean (SD)	Year-4 Assessment Mean (SD)
Psychometric Intelligence			
Full Scale IQ	104.1 (7.4)	106.5 (11.0)	110.7 (10.4)
Verbal IQ	100.1 (9.9)	103.4 (10.3)	105.3 (10.5)
Performance IQ	107.9 (9.6)	108.8 (12.7)	114.5 (12.6)
Information Scaled Score	9.1 (2.6)	9.4 (2.3)	9.4 (2.5)
Comprehension Scaled Score	10.4 (3.0)	10.2 (2.5)	10.6 (2.6)
Digit Span Scaled Score	9.6 (2.3)	9.8 (3.1)	11.1 (3.1)
Arithmetic Scaled Score	10.9 (2.6)	11.0 (2.5)	10.8 (2.5)
Similarities Scaled Score	10.3 (3.0)	11.3 (2.6)	12.0 (2.2)
Vocabulary Scaled Score	9.5 (2.2)	10.9 (2.3)	10.8 (2.4)
Picture Completion Scaled Score	12.2 (2.4)	10.7 (3.1)	12.3 (3.1)
Picture Arrangement Scaled Score	11.0 (2.4)	10.9 (2.7)	11.3 (2.6)

Table 12 (Continued)

	Year-0 Assessment Mean (SD)	Year-2 Assessment Mean (SD)	Year-4 Assessment Mean (SD)
Block Design Scaled Score	11.2 (2.8)	12.1 (2.8)	12.5 (3.3)
Object Assembly Scaled Score	11.2 (2.7)	11.4 (2.9)	12.9 (3.4)
Coding Scaled Score	10.1 (2.6)	11.2 (2.4)	11.7 (2.4)
<u>Academic Achievement and Reading</u>			
MAT Word Knowledge Standard Score	42.3 (9.6)	51.0 (9.7)	51.6 (9.4)
MAT Word Discrimination Standard Score	47.3 (10.5)	51.2 (10.1)	--- *
MAT Reading Standard Score	40.7 (10.0)	46.6 (10.4)	50.1 (9.8)
WRAT Reading Standard Score	104.6 (17.5)	109.2 (19.9)	111.0 (19.6)
WRAT Spelling Standard Score	102.6 (14.9)	102.3 (16.6)	103.8 (19.4)
WRAT Arithmetic Standard Score	97.6 (6.0)	97.6 (8.9)	97.5 (10.6)
<u>Auditory-Perception and Language</u>			
PPVT-IQ	106.4 (12.4)	107.9 (11.7)	105.9 (9.4)

Table 12 (Continued)

	Year-0 Assessment Mean (SD)	Year-2 Assessment Mean (SD)	Year-4 Assessment Mean (SD)
Auditory Closure correct	11.4 (4.4)	14.9 (3.7)	19.4 (3.4)
Sentence Memory correct	11.5 (2.5)	13.4 (2.5)	16.1 (1.9)
Verbal Fluency correct	5.4 (2.6)	7.9 (2.2)	10.3 (3.1)
Auditory (R) errors	.1 (.2)	-- *	-- *
Auditory (L) errors	.1 (.7)	.1 (.4)	.02 (.2)
Speech-sounds Perception correct	19.3 (6.1)	24.8 (3.2)	27.1 (2.1)
Motor			
Finger Tapping (R) taps/10"	29.1 (5.2)	33.2 (5.1)	39.0 (5.0)
Finger Tapping (L) taps/10"	27.0 (4.5)	31.0 (4.3)	36.5 (4.9)
Foot Tapping (R) taps/10"	25.4 (5.0)	30.2 (4.4)	36.3 (4.7)
Foot Tapping (L) taps/10"	23.7 (4.1)	28.1 (4.3)	34.2 (4.7)
Maze Speed (R)	111.4 (31.1)	107.7 (28.4)	102.9 (35.8)
Maze Speed (L)	111.3 (31.5)	100.7 (22.4)	98.3 (29.0)

Table 12 (Continued)

	Year-0 Assessment Mean (SD)	Year-2 Assessment Mean (SD)	Year-4 Assessment Mean (SD)
Maze Time (R)	4.5 (3.3)	2.0 (1.4)	1.8 (1.8)
Maze Time (L)	9.3 (5.1)	5.9 (3.1)	4.2 (2.9)
Maze Counter (R) errors	31.6 (17.1)	16.0 (10.5)	12.8 (10.0)
Maze Counter (L) errors	61.2 (28.1)	42.6 (18.0)	28.8 (15.6)
Grip Strength (R) Kg.	12.4 (2.4)	13.6 (3.0)	19.6 (5.0)
Grip Strength (L) Kg.	11.8 (2.2)	12.3 (2.8)	17.8 (3.5)
Name Writing speed (R)	19.0 (6.3)	12.2 (5.2)	8.0 (2.2)
Name Writing speed (L)	35.3 (12.7)	25.6 (9.2)	17.3 (5.1)
Visual-Perception			
Target correct	13.7 (3.0)	16.7 (2.6)	18.5 (2.2)
Rhymes correct	14.8 (4.7)	18.0 (3.7)	18.9 (1.2)
Reverses correct	32.7 (4.4)	39.0 (4.2)	40.7 (3.0)
Visual (R) errors	.1 (.4)	.2 (.5)	.02 (.2)
Visual (L) errors	.2 (.5)	.1 (.4)	.2 (.6)

Table 12 (Continued)

	Year-0 Assessment Mean	(SD)	Year-2 Assessment Mean	(SD)	Year-4 Assessment Mean	(SD)
Tactile-Perception						
Finger Agnosia (R) errors	1.8	(1.8)	.7	(1.1)	.5	(1.1)
Finger Agnosia (L) errors	2.2	(2.2)	.7	(1.0)	.4	(1.0)
Tactile (R) errors	1.0	(1.2)	.3	(.8)	.2	(.4)
Tactile (L) errors	.7	(1.1)	.2	(.7)	.2	(.6)
Underlining Test						
Subtest 1	26.0	(7.4)	23.4	(5.4)	30.3	(6.4)
Subtest 2	28.4	(7.6)	25.2	(4.5)	32.2	(5.3)
Subtest 3	17.1	(5.0)	14.8	(4.2)	20.4	(3.3)
Subtest 4	7.3	(5.4)	7.2	(4.4)	12.0	(5.3)
Subtest 5	21.9	(6.5)	19.3	(4.0)	24.1	(4.5)
Subtest 6	15.3	(5.3)	14.4	(3.0)	19.6	(5.2)
Subtest 7	15.4	(4.8)	12.7	(3.1)	16.1	(3.5)
Subtest 8	8.5	(3.8)	7.1	(1.2)	9.2	(2.4)

Table 12 (Continued)

	Year-0 Assessment Mean (SD)	Year-2 Assessment Mean (SD)	Year-4 Assessment Mean (SD)
Subtest 9	8.0 (3.4)	6.3 (2.1)	9.0 (2.5)
Subtest 10	12.2 (6.2)	11.5 (4.4)	16.6 (4.1)
Subtest 11	11.9 (6.7)	13.8 (4.1)	17.7 (3.2)
Subtest 12	7.0 (3.2)	6.7 (1.8)	10.1 (2.3)
Subtest 13	23.8 (7.6)	20.5 (4.3)	25.7 (5.8)
<u>Right-Left Awareness</u>			
Right-Left Awareness correct	17.9 (4.8)	19.9 (4.7)	22.5 (4.0)

* missing data

.85 (WRAT Reading subtest) to .28 (Picture Completion subtest). The mean Pearson-r coefficient calculated across all 63 variables was .37 (SD = .24). A coefficient could not be calculated for the Auditory (R) errors measure because of insufficient data. The rank order of these Pearson-r coefficients is provided in Table 13. All of the correlation coefficients calculated for the Academic Achievement and Reading measures were statistically significant. The WRAT Reading subtest provided the greatest measure of consistency of all variables (r = .85). The WRAT Arithmetic subtest was associated with the least coefficient with regards to magnitude (r = .51). Regarding the consistency of the measures within this ability domain, the ranking of the variables ranged from 1 (WRAT Reading subtest) to 18 (WRAT Arithmetic subtest). The median rank was 4 for these measures.

All of the measures of Auditory-Perception and Language provided statistically significant consistency coefficients. The greatest consistency was observed for the Auditory (L) errors variable (r = .83), while Verbal Fluency correct (r = .39) demonstrated the least degree of consistency within this ability area. Ranks ranged from 5 for the former measure, to 30th for the latter. The ability domain was associated with a median rank value of 16.

Twelve of the 14 measures of Psychometric Intelligence yielded significant Pearson-r correlation coefficients.

Table 13

Rank Order of Year-0 to Year-2 Pooled Subjects' Tests Based on Magnitude of Correlation (Pearson-r)

Rank	Variable	Pearson-r
1	WRAT Reading Standard Score	.85
2	WRAT Spelling Standard Score	.84
3	Auditory (L) errors	.83
4	MAT Word Discrimination Standard Score	.77
5	MAT Reading Standard Score	.76
6	Full Scale IQ	.70
7	Grip Strength (R) Kg.	.69
8	Rhymes correct	.68
9	MAT Word Knowledge Standard Score	.64
10	Grip Strength (L) Kg.	.62
11	Auditory Closure correct	.61
12	Verbal IQ	.60
13	Information Scaled Score	.59
14	Performance IQ	.55
15	Target correct	.55
16	Sentence Memory correct	.53
17	PPVT-IQ	.51
18	WRAT Arithmetic Standard Score	.51
19	Foot Tapping (L) taps/10"	.50
20	Digit Span Scaled Score	.49
21	Speech-sounds Perception correct	.47
22	Similarities Scaled Score	.44
23	Foot Tapping (R) taps/10"	.44
24	Tactile (L) errors	.43
25	Maze Counter (L) errors	.42

Table 13 (Continued)

Rank	Variable	Pearson-r
26	Block Design Scaled Score	.41
27	Maze Counter (R) errors	.41
28	Finger Tapping (R) taps/10"	.40
29	Maze Speed (R)	.39
30	Verbal Fluency correct	.39
31	Vocabulary Scaled Score	.38
32	Finger Tapping (L) taps/10"	.38
33	Object Assembly Scaled Score	.38
34	Maze Time (L)	.38
35	Picture Arrangement Scaled Score	.37
36	Arithmetic Scaled Score	.37
37	Underlining Subtest 12	.35
38	Maze Speed (L)	.33
39	Underlining Subtest 11	.32
40	Maze Time (R)	.31
41	Picture Completion Scaled Score	.28
42	Underlining Subtest 10	.27*
43	Right-Left Awareness correct	.26*
44	Coding Scaled Score	.26*
45	Reverse correct	.26*
46	Underlining Subtest 4	.24*
47	Comprehension Scaled Score	.21*
48	Underlining Subtest 6	.21*
49	Name Writing speed (R)	.19*
50	Finger Agnosia (R) errors	.13*
51	Name Writing speed (L)	.11*
52	Underlining Subtest 3	.09*
53	Underlining Subtest 13	.09*
54	Finger Agnosia (L) errors	.09*
55	Underlining Subtest 7	.06*

Table 13 (Continued)

Rank	Variable	Pearson-r
56	Underlining Subtest 8	.03*
57	Visual (L) errors	.01*
58	Underlining Subtest 9	.004*
59	Tactile (R) errors	-.01*
60	Underlining Subtest 2	-.03*
61	Visual (R) errors	-.03*
62	Underlining Subtest 5	-.06*
63	Underlining Subtest 1	-.08*
64	Auditory (R) errors	--++

* $p > .05$

++ Insufficient data to calculate Pearson-r coefficient

These measures ranged in the magnitude of their coefficients from .70 (WISC Full Scale IQ) to .28 (Picture Completion subtest). The rankings of these variables ranged from 6 for the latter measure to 47 for the Comprehension subtest. The median Psychometric Intelligence rank value was calculated to be 28.

Twelve of the 14 measures of Motor ability produced significant Pearson-*r* coefficients. Regarding the consistency of these measures, Grip Strength (R) revealed the greatest relationship (*r* = .69) and was rated 7th in the overall rank order distribution. The least consistent of these 12 variables was Maze Time (R) (*r* = .31). The lowest ranking of all the Motor variables was evidenced by the Name Writing speed (L) measure, which was rated 51st in terms of consistency. The overall median rank value was found to be 28.

For the 13 subtests of the Underlining Test, two revealed significant correlation coefficients. The greatest magnitude in correlation coefficient was .35 (Subtest 12). The range of ranks, related to retest consistency, was from 37 (Subtest 12) to 63 (Subtest 1). The median value for all the measures within the Underlining Test category was 53.

Only the Tactile (L) errors measure of Tactile-Perceptual ability (*r* = .43), and the Rhymes correct and Target correct measures of Visual-Perceptual ability (*r* = .68 and .55, respectively) were significant Pearson-*r*

coefficients. The measure of Right-Left Awareness failed to provide a statistically significant correlation coefficient (r = .26). The median rankings for these four ability areas were: 24, 8, 15, and 51, in order.

Considering the stability of the neuropsychological measures over the initial retest interval, the ICC coefficients for each ability domain were examined. The total range of ICC coefficient values was from .76 (Auditory (L) errors) to -.17 (Arithmetic subtest). Thirty of the 64 measures provided statistically significant ($p < .05$) ICC coefficients. The mean ICC coefficient calculated for these measures was found to be .19 (SD = .17). The rank order of the ICC coefficients associated with the test variables, calculated over the initial retest interval, is provided in Table 14.

Nine of the 63 measures were associated with negative ICC coefficient values. Of these nine measures, five (Tactile (R) errors, Name Writing speed (R), Name Writing speed (L), Maze Time (L), and Visual (R) errors) were indicative of a general improvement in the pooled sample's performance. Of the remaining 54 measures, 10 variables (Tactile (L) errors, Finger Agnosia (R) errors, Finger Agnosia (L) errors, Maze Speed (R), Maze Speed (L), Maze Counter (R) errors, Maze Counter (L) errors, Maze Time (R), Visual (L) errors, and Auditory (L) errors indicated a general worsening of childrens' performance. An ICC value

Table 14

Rank Order of Year-0 to Year-2 Pooled Subjects' Tests Based on Magnitude of Correlation (ICC)

Rank	Variable	ICC
1	Auditory (L) errors	.76
2	MAT Reading Standard Score	.49
3	MAT Word Knowledge Standard Score	.47
4	MAT Word Discrimination Standard Score	.45
5	Grip Strength (R) Kg.	.45
6	WRAT Reading Standard Score	.42
7	Finger Tapping (R) taps/10"	.38
8	Vocabulary Scaled Score	.35
9	Object Assembly Scaled Score	.34
10	Target correct	.34
11	Grip Strength (L) Kg.	.34
12	Tactile (L) errors	.33
13	Performance IQ	.32
14	WRAT Spelling Standard Score	.32
15	Rhymes correct	.31
16	Finger Tapping (L) taps/10"	.31
17	Sentence Memory correct	.31
18	Underlining Subtest 12	.30
19	Underlining Subtest 11	.29
20	WRAT Arithmetic Standard Score	.29
21	Verbal Fluency correct	.28
22	Verbal IQ	.28
23	Information Scaled Score	.26
24	Auditory Closure correct	.26
25	Underlining Subtest 10	.25

Table 14 (Continued)

Rank	Variable	ICC
26	Digit Span Scaled Score	.25
27	Full Scale IQ	.25
28	Foot Tapping (R) taps/10"	.25
29	Underlining Subtest 4	.24
30	Foot Tapping (L) taps/10"	.23
31	Maze Speed (R)	.21*
32	Maze Counter (R)	.21*
33	Underlining Subtest 6	.18*
34	PPVT-IQ	.17*
35	Reverse correct	.16*
36	Similarities Scaled Score	.16*
37	Picture Completion Scaled Score	.14*
38	Picture Arrangement Scaled Score	.14*
39	Block Design Scaled Score	.14*
40	Right-Left Awareness correct	.14*
41	Maze Speed (L)	.14*
42	Underlining Subtest 3	.10*
43	Visual (L) errors	.10*
44	Underlining Subtest 13	.08*
45	Comprehension Scaled Score	.07*
46	Speech-sounds Perception correct	.07*
47	Underlining Subtest 7	.06*
48	Finger Agnosia (R) errors	.05*
49	Maze Counter (L) errors	.05*
50	Coding Scaled Score	.05*
51	Maze Time (R)	.04*
52	Underlining Subtest 8	.02*
53	Underlining Subtest 9	.004*
54	Finger Agnosia (L) errors	.001*

Table 14 (Continued)

Rank	Variable	ICC
55	Tactile (L) errors	-.02*
56	Underlining Subtest 2	-.03*
57	Name Writing speed (R)	-.03*
58	Visual (R) errors	-.03*
59	Underlining Subtest 5	-.05*
60	Underlining Subtest 1	-.08*
61	Maze Time (L)	-.08*
62	Name Writing speed (L)	-.11*
63	Arithmetic Scaled Score	-.17*
64	Auditory (R) errors	-- ++

* $p > .05$

++ Insufficient data to calculate ICC coefficient

for the Auditory (R) errors measure could not be calculated due to a lack of data.

The Academic Achievement and Reading measures demonstrated very little variation in terms of their ICC values. Three of the four measures with the greatest stability belong to this ability area (MAT Reading, Word Knowledge, and Word Description subtests; $\underline{r}$ = .49, .47, and .45, respectively). The WRAT Arithmetic subtest had the lowest ICC coefficient associated with it ($\underline{r}$ = .29). The rankings for these six Academic Achievement and Reading measures were generally well up into the upper end of the rank order distribution, with the median rank equalling 5.

Six Auditory-Perception and Language variables were involved in the calculation of ICC values. Stability values ranged from .76 (Auditory (L) errors) to .07 (Speech-sounds Perception correct). Rankings for these measures were 1st and 46th, with respect to stability. The median stability rank for this ability domain was 22.

The stability values associated with the Psychometric Intelligence measures ranged from .35 (Vocabulary subtest) to -.17 (Arithmetic subtest). The median rank value for these 14 measures was 31, and individual measures' ranks ranged from 8 (Vocabulary subtest) to 63 (Arithmetic subtest).

A large degree of variation was observed for the ICC values associated with the Motor ability measures. The

greatest stability was seen for the measure of Grip Strength (R) ($\underline{r}$ = .45), while the least stability was associated with Name Writing speed (L) ($\underline{r}$ = -.11). The rankings of these two measures, with respect to the stability of all the measures, was 5th and 62, in order. The median stability rank value for the Motor ability domain was 31.

The 13 measures from the Underlining Test were associated with ICC values ranging from .30 (Subtest 12) to -.08 (Subtest 1). The range of stability rankings for these measures was equally variable. Subtest 11 was associated with a rank value of 18, while Subtest 1 was ranked 60th. The median rank for the entire ability realm was 44.

The five Visual-Perceptual measures ranged in terms of their ICC coefficient values from .34 (Target correct) to -.03 (Visual (R) errors). The rankings of these measures was from 10th to 58th, in regards to their relative stability. The median stability rank value for the Visual-Perceptual variables was calculated to equal 34.

The four Tactile-Perceptual variables were generally associated with low ICC coefficient values. The Tactile (L) errors measure had the greatest stability coefficient ($\underline{r}$ = .33), and was ranked 12th in terms of its relative stability. The remaining three measures had coefficients ranging from .05 (Finger Agnosia (R) errors)to -.02 (Tactile (R) errors). The range of ranked values for these measures was from 48 to 55, respectively. The median rank value for

this ability area was 51.

Finally, the sole measure of Right-Left Awareness was associated with an ICC coefficient value of .14, and was afforded a relative stability rank of 38.

Year-2/Year-4 Two-Year Retest Interval

Of the 61 Pearson-r coefficients that compare the same subjects' performance at the Year-2 assessment with that at the Year-4 assessment, only 12 measures proved to be statistically significant to the $p < .05$ level. The range of the Pearson-r coefficients extended from .56 (MAT Reading subtest) to -.17 (Finger tapping (L)). The average coefficient value for this assessment period is .13 (SD = .18). Coefficients could not be calculated for three measures (Auditory (L) errors, Auditory (R) errors, and MAT Word Discrimination subtest) due to insufficient data. Table 15 contains the Pearson-r coefficients for all 61 variables, ranked in order of magnitude.

Four of the five Academic Achievement and Reading measures provided significant Pearson-r coefficients. The range of these coefficients was from .56 (MAT Reading subtest) to .20 (WRAT Arithmetic subtest). All four of the significant measures occupied ranks within the top 10 placings for consistency. The median rank for the Academic Achievement and Reading measures was 3.

The Pearson-r coefficient associated with the Right-Left Awareness measure ($r = .36$) was statistically

Table 15

Rank Order of Year-2 to Year-4 Pooled Subjects' Tests Based on Magnitude of Correlation (Pearson-r)

Rank	Variable	Pearson-r
1	MAT Reading Standard Score	.56
2	MAT Word Knowledge Standard Score	.52
3	WRAT Reading Standard Score	.49
4	Verbal IQ	.46
5	Verbal Fluency correct	.41
6	Underlining Subtest 8	.36
7	Right-Left Awareness correct	.36
8	Arithmetic Scaled Score	.36
9	WRAT Spelling Standard Score	.34
10	Full Scale IQ	.33
11	Information Scaled Score	.32
12	Rhymes correct	.32
13	Underlining Subtest 12	.29*
14	Sentence Memory correct	.29*
15	Underlining Subtest 6	.27*
16	Foot Tapping (L) taps/10"	.27*
17	Digit Span Scaled Score	.25*
18	Grip Strength (R) Kg.	.23*
19	Vocabulary Scaled Score	.23*
20	PPVT-IQ	.22*
21	WRAT Arithmetic Standard Score	.20*
22	Auditory Closure correct	.20*
23	Similarities Scaled Score	.20*
24	Reverse correct	.17*
25	Object Assembly Scaled Score	.14*
26	Block Design Scaled Score	.13*
27	Target correct	.12*

Table 15 (Continued)

Rank	Variable	Pearson-r
28	Visual (L) errors	.12*
29	Underlining Subtest 7	.11*
30	Maze Speed (R)	.11*
31	Tactile (R) errors	.10*
32	Performance IQ	.09*
33	Underlining Subtest 11	.08*
34	Foot Tapping (R) taps/10"	.08*
35	Underlining Subtest 4	.07*
36	Picture Arrangement Scaled Score	.07*
37	Coding Scaled Score	.07*
38	Finger Tapping (R) taps/10"	.07*
39	Underlining Subtest 3	.06*
40	Speech-sounds Perception correct	.04*
41	Underlining Subtest 1	.03*
42	Finger Agnosia (L) errors	.02*
43	Comprehension Scaled Score	.02*
44	Maze Speed (L)	.02*
45	Maze Counter (L) errors	-.03*
46	Name Writing speed (R)	-.04*
47	Grip Strength (L) Kg.	-.05*
48	Visual (R) errors	-.05*
49	Name Writing speed (L)	-.05*
50	Tactile (L) errors	-.06*
51	Underlining Subtest 7	-.06*
52	Underlining Subtest 10	-.07*
53	Maze Time (L)	-.07*
54	Maze Time (R)	-.08*
55	Maze Counter (R) errors	-.10*
56	Picture Completion	-.10*
57	Underlining Subtest 9	-.11*

Table 15 (Continued)

Rank	Variable	Pearson-r
58	Underlining Subtest 5	-.13*
59	Finger Agnosia (R) errors	-.13*
60	Underlining Subtest 13	-.14*
61	Finger Tapping (L) taps/10"	-.17*
62	Auditory (L) errors	--- ++
63	MAT Word Discrimination Standard Score	--- ++
64	Auditory (R) errors	--- ++

* $p > .05$

++ Insufficient data to calculate Pearson-r coefficient

significant. With respect to consistency, this variable was ranked 7th.

Only the Verbal Fluency correct measure of the Auditory-Perception and Language ability domain was associated with a significant Pearson-r coefficient. The range of the coefficient values within this domain was from .41 (Verbal Fluency correct) to .04 (Speech-sounds Perception correct). All of the measures were ranked greater than 41st in terms of their consistency. The range of ranked values was 5th (Verbal Fluency correct) to 40th (Speech-sounds Perception correct). The median consistency rank value for this ability is 17.

Four measures of Psychometric Intelligence provided significant Pearson-r coefficients. The range of the coefficients was from .46 (WISC Verbal IQ) to -.10 (Picture Completion subtest). Nine of the 14 measures were ranked within the upper half of the rank order distribution relative to their consistency. The median Psychometric Intelligence rank value was 24.

One measure of Visual-Perception, Rhymes correct, was associated with a significant correlation coefficient. The magnitude of the coefficient for this variable was .32, and it was ranked 12 with respect to its relative consistency. The least consistent variable within this domain was the Visual (R) errors measure ($r = -.05$), which was ranked 48th. The median value for all Visual-Perceptual measures was 27.

Subtest 8 of the Underlining Test provided the only significant correlation coefficient within this domain. Associated with a coefficient of .36, this subtest was ranked 6th overall for consistency. The least consistent measure within the Underlining Test subtests was Subtest 13 ($\underline{r}$ = -.14). The majority of the subtests were ranked within the lower half of the distribution. A rank of 39 was the median value for the Underling Test measures.

None of the Motor ability measures yielded a significant correlation coefficient. The greatest consistency was demonstrated by the Foot Tapping (L) measures, with a coefficient of .27. The least consistent variable was the Finger Tapping (L) measure, which had a coefficient of -.17 associated with it. Nine of the 14 variables received consistency rankings within the lower one-third of the distribution. Median rank value for these tests was 46.

Finally, none of the Tactile-Perceptual measures' Pearson-$\underline{r}$ coefficients proved to be of any significant magnitude. The range of these measures was from .10 (Tactile (R) errors) to -.13 (Finger Agnosia (R) errors). All of these measures were ranked well below the other ability domains in terms of their consistency. The range of the ranks was 31st for Tactile (R) errors, to 59th for Finger Agnosia (R). The median value for these variables was 46.

The ICC correlation coefficients comparing the childrens' performance on the neuropsychological variables at Year-2 and Year-4 are provided, in ranked format, in Table 16. The magnitude of the ICC coefficients ranged from .49 (Visual (L) errors) to -.19 (MAT Word Knowledge subtest). Only seven of these measures were significant to the $p < .05$ level. The mean ICC coefficient calculated for these variables was .04 (SD = .16). Coefficients could not be calculated for three measures (Auditory (L) errors, Auditory (R) errors, and MAT Word Description subtest) due to insufficient data. Fully half of the 61 measures examined were associated with negative ICC coefficients. Of the 30 measures associated with negative ICC coefficient values, five (Visual (R) errors, Maze Time (R), Maze Speed (R), Maze Speed (L), and, Tactile (L) errors) were indicative of an improvement in the childrens' performance. Eight of the 31 measures with position ICC values (Finger Agnosia (L) errors, Finger Agnosia (R) errors, Tactile (R) errors, Name Writing speed (R), Name Writing speed (L), Maze Time (L), Maze Counter (L) errors, and Maze Counter (R) errors) indicated that a general worsening of childrens' performances on these variables.

Relative to stability, three of four Tactile-Perceptual measures were ranked 3rd (Finger Agnosia (L) errors) to 12th (Tactile (R) errors). The ICC values for these three variables ranged from .37 to .20, respectively. The median

Table 16

Rank Order of Year-2 to Year-4 Pooled Subjects' Tests Based on Magnitude of Correlation (ICC)

Rank	Variable	ICC
1	Visual (L) errors	.49
2	Coding Scaled Score	.37
3	Finger Agnosia (L) errors	.37
4	Maze Time (L)	.31
5	Underlining Subtest 8	.30
6	Underlining Subtest 12	.28
7	Finger Agnosia (R) errors	.27
8	Name Writing speed (R)	.23*
9	Name Writing speed (L)	.23*
10	Underlining Subtest 6	.22*
11	Finger Tapping (R) taps/10"	.21*
12	Tactile (R) errors	.20*
13	Object Assembly Scaled Score	.19*
14	Rhymes correct	.17*
15	Digit Span Scaled Score	.15*
16	Finger Tapping (L) taps/10"	.13*
17	Arithmetic Scaled Score	.12*
18	Speech-sounds Perception correct	.12*
19	WRAT Arithmetic Scaled Score	.11*
20	Underlining Subtest 2	.10*
21	Underlining Subtest 11	.09*
22	Target correct	.08*
23	Verbal IQ	.08*
24	Underlining Subtest 4	.07*
25	Underlining Subtest 3	.06*
26	Maze Counter (L) errors	.06*

Table 16 (Continued)

Rank	Variable	ICC
27	Picture Completion Scaled Score	.06*
28	Full Scale IQ	.03*
29	Underlining Subtest 1	.03*
30	Maze Counter (R) errors	.01*
31	Auditory Closure correct	.01*
32	Right-Left Awareness correct	-.003*
33	Visual (R) errors	-.02*
34	WRAT Reading Standard Score	-.02*
35	Comprehension Scaled Score	-.02*
36	Performance IQ	-.03*
37	Maze Time (R)	-.03*
38	Sentence Memory correct	-.03*
39	Picture Arrangement Scaled Score	-.05*
40	Underlining Subtest 7	-.06*
41	Underlining Subtest 10	-.07*
42	Tactile (L) errors	-.08*
43	Verbal Fluency correct	-.08*
44	Foot Tapping (R) taps/10"	-.09*
45	MAT Reading Standard Score	-.10*
46	Underlining Subtest 9	-.11*
47	WRAT Spelling Standard Score	-.11*
48	Information Scaled Score	-.12*
49	Grip Strength (R) Kg.	-.12*
50	Foot Tapping (L) taps/10"	-.12*
51	Grip Strength (L) Kg.	-.12*
52	PPVT-IQ	-.12*
53	Underlining Subtest 13	-.12*
54	Underlining Subtest 5	-.13*
55	Similarities Scaled Score	-.13*
56	Maze Speed (R)	-.13*

Table 16 (Continued)

Rank	Variable	ICC
57	Vocabulary Scaled Score	-.14*
58	Reverse correct	-.15*
59	Block Design Scaled Score	-.17*
60	Maze Speed (L)	-.17*
61	MAT Word Knowledge Standard Score	-.19*
62	Auditory (L) errors	-- ++
63	Auditory (R) errors	-- ++
64	MAT Word Discrimination Standard Score	-- ++

* $p > .05$

++ Insufficient data to calculate ICC coefficient

rank value for the Visual-Perceptual measures was 9.

The five Visual-Perceptual measures were very variable in terms of their ICC coefficient magnitudes. A high value of .49 was observed for the Visual (L) errors measure (stability rank of 1), and a low ICC value of -.15 for Reversals correct (stability rank of 58th). The Visual-Perceptual median value was 22nd.

The Underlining Test measures were associated with ICC values ranging from .30 (Subtest 8) to -.13 (Subtest 5). The range of the ranks assigned to these variables was from 5th to 54th. The median rank value for the Underlining measures was 25.

The 14 Psychometric Intelligence measures had ICC values ranging from .37 (Coding subtest) to -.17 (Block Design subtest) associated with them. The stability ranking associated with these two measures were 2 and 59, respectively. the median stability ranking for this ability was 31.

The measure of Right-Left Awareness was observed to have an ICC value of -.003, and was ranked 32nd in terms of its stability.

The range of the ICC values for the 14 Motor variables closely resembled that observed for the Psychometric Intelligence measures. An ICC coefficient of .31 was associated with the Maze Time (L) measure, while the least stable measure within this domain was Maze Speed (L) ($\underline{r}$ = -

.17). The rankings of the relative stability was 4th to 60th for these measures, respectively. Median rank for all Motor variables was 34.

The five Auditory-Perception and Language measures ranged in terms of their ICC values from .12 (Speech-Sounds Perception correct) to -.12 (PPVT-IQ). Stability rankings for these same variables was from 18th to 52nd, respectively. The median stability rank for all Auditory-Perception and Language measures was .40.

Lastly, the five Academic Achievement and Reading measures had generally low ICC values associated with them. The greatest magnitude of ICC coefficient was for the WRAT Arithmetic subtest score (r = .11) and received a stability ranking of 15. The MAT Word Knowledge subtest was ranked last (61st) on the basis of its ICC coefficient value of -.19. Median rank for these five measures was 46.

Comparing the rank order of the neuropsychological variables' consistency between the first and second two-year assessment intervals, a Spearman rank-order correlation coefficient was generated using the NONPAR CORR SPSSx program (Nie, 1983). The resulting coefficient proved to be highly significant ($r(62) = .43$; $p < .001$). A similar comparison between the two rank orders of ICC coefficient yielded a significant, but negative correlation coefficient, ($r(59) = -.26$, $p < .05$).

The above investigations had four specific goals to

accomplish. First, the two-year test-retest stability and consistency of the DR childrens' performance on a variety of neuropsychological variables was assessed. Second, the degree and pattern of consistency and stability found for the DR children's performance was compared to that found for the NR children. In this manner, the degree of similarity or differences between the two samples' performance could be investigated. Third, additional comparisons were made between the DR children and two clinical samples to determine the degree of relatedness between the DR childrens' patterns of stability and consistency with those found for the two clinical samples. Fourth, a partial validation was attempted by comparing the two-year retest stability and consistency at one assessment period with that for a second, equal assessment period. In this latter study, all subjects were pooled together to form a single group. A discussion of the results follows.

CHAPTER IV

DISCUSSION

The order of presentation for this discussion parallels that used previously in the Results section. First, the two-year retest consistency and stability of the DR subjects' performance on a variety of neuropsychological measures is discussed, followed by a similar discussion of the NR subjects' performance; this allows for a comparison of the patterns of stability and consistency of the two groups. Second, the retest performance of the DR children is compared to that for two heterogeneous clinical samples in order to elicit any similarities or differences between the two types of children. Lastly the pattern of consistency and stability in the performance of the total sample over several neuropsychological measures is discussed.

The interpretation and discussion of the resulting correlation coefficients found in Investigations 1 through 4 is made utilizing the standards for predictive value presented by Brown et al. (1989). The values of the Pearson-r and ICC coefficients are rated for their predictive value on the basis of the magnitude of the coefficient. Coefficients equalling or exceeding a value of

.60 are rated as having "excellent" consistency or stability (as the case may be). Coefficient values ranging upwards from .40 to .59 are considered to be "good" predictors. A range of values from .30 to .39 is associated with "fair" prediction. Finally, Pearson-r or ICC coefficients that fail to meet to exceed a value of .30 are be rated as poor predictors for this presentation.

Investigations 1 and 2

The first part of this section is directed towards discussing the results that emerged from Investigations 1 and 2. Examination of the consistency of the 25 DR childrens' performance on the neuropsychological measures over a two-year retest interval reveals that the majority of the measures are rated as poor predictors of performance. However, of the 21 variables found to be associated with statistically significant Pearson-r coefficients, most are good or excellent measures of consistency.

Considering separately the consistency of the variables that comprise each of the eight ability areas (Psychometric Intelligence, Academic Achievement and Reading, Auditory-Perception and Language, Motor, Visual-Perceptual, Tactile-Perceptual, the Underlining Test measures, and Right-Left Awareness) it appears that in general, the performances of the DR children are most consistent on those variables that fall within the Motor domain. The majority of these measures are rated to have good to excellent consistency.

A lesser degree of consistency is observed for the Psychometric Intelligence, Academic Achievement and Reading, and Auditory-Perception and Language measures. The sample's retest performance on the most of these variables is rated as being fair or better in consistency. However, the median rankings associated with each ability category suggests that, while their performance is less consistent within these three ability areas (relative to Motor ability) it remains more consistent than that observed for the Visual-Perceptual, Tactile-Perceptual, Underlining Test measures, and Right-Left Awareness measures. This is thought to be the case despite a large degree of discrepancy between four of the five Visual-Perception measures that may be masked by the median rank value. Figure 6 illustrates this pattern of reliability. The DR sample demonstrates excellent consistency in their performance on two of the Visual-Perception measures (Rhymes correct and Target correct). In contrast, they show little consistency on the Visual-Perceptual measures of Visual (R) errors and Visual (L) errors.

Finally, the Tactile-Perceptual and Right-Left Awareness ability domains, as well as the Underlining Test measures, were generally associated with the smallest degree of retest consistency. None of these measures yielded statistically significant correlation coefficients.

The performances of the DR children over the retest

interval is generally poor. Only four variables (Target correct, Finger Tapping (R), Maze Speed (L), and Rhymes correct) were associated with ICC coefficients of sufficient magnitude to be rated as good in stability.

Between the eight ability domains, the childrens' scores on the Psychometric Intelligence measures are the most stable, relatively speaking, followed by the Academic Achievement and Reading and Visual-Perceptual variables. The stability of the measures encompassed by the Motor and Auditory-Perceptual and Language abilities appear to be less than that demonstrated for the previously mentioned ability domains. Finally, the least stable performance over time is evident for the Tactile-Perceptual domain, Underlining Test, and Right-Left Awareness measures. It is important to recognize, that these are just relative differences between ability domains. In all but four cases the DR childrens' performance is shown to have poor stability.

The pattern of consistency associated with the 27 NR childrens' two-year retest performance on the neuropsychological measures differs from that just discussed for the DR children. This conclusion is suggested by the non-significant rank order correlation coefficient used to compare the two groups' rankings of Pearson-r coefficients. The NR children demonstrate good or excellent consistency in their assessments on 32 variables, as compared to only 16 found for the DR children.

The two samples' different patterns of consistency between ability areas is evident. Such differences between the patterns is readily apparent in Figures 1 through 7. The NR children reveal good to excellent consistency in their performance on a majority of the measures contained within the Academic Achievement and Reading, and Psychometric Intelligence domains. In contrast, the DR children are more variable in their performance on these measures. The most consistent (although only fair) performance for the DR children is on tests that fall within the Motor domain.

There are some similarities, however, in the two samples' consistency in their re-assessment on measures of Visual-Perception and Tactile-Perception. Both groups generally evidence poor consistency in their performance on measures that are often interpreted from the perspective of pathognomonic signs rather than levels of performance (e.g., Visual (R) errors, Finger Agnosia (L), and Tactile (R) errors.

Diferences are also evident between the groups' stability of performance over time. The most salient difference between the NR and DR childrens' two-year retest performance on the neuropsychological variables is the generally lesser degree of stability in evidence for the DR sample's assessments over time. This finding is reflected in the greater number of the NR childrens' good to excellent

retest performance on measures, relative to that for the DR children, and, in their greater mean ICC coefficient value than that yielded by the DR sample.

Examining the relative rankings of each sample's performance in the eight ability areas, with respect to stability, some differences are evident. The NR childrens' performance is the most stable for the Academic Achievement and Reading measures, followed by the Psychometric Intelligence tests. The NR sample's assessment on the majority of these measures is found to have good to excellent stability. For the DR children's performance, however, the greatest stability is observed on the Psychometric Intelligence variables, followed by measures falling within the Visual-Perceptual, Academic Achievement and Reading, and Auditory-Perception and Language ability domains. In marked contrast to the NR group's performance, however, virtually all of the measures within those ability areas are rated as poor.

The performance of the NR children is associated with poor stability on the Tactile-Perceptual, Visual-Perceptual, Right-Left Awareness, and Underlining Test measures. While similar ratings of stability are afforded to the DR group's performance on these same variables, those contained within the Visual-Perceptual category rank among the most stable of all the measures (second only to the Psychometric Intelligence measures) for these children.

To summarize the retest consistency and stability of both groups' performance on the neuropsychological variables over a two-year interval, several findings are in evidence. The performances by the NR group are generally more consistent and stable than those seen for the DR children. The pattern of consistency and stability associated with the DR children is statistically different from that observed for the NR children. The degree of stability and consistency of the NR group's performance on the two test administrations is found to be greatest for the Motor variables. A lesser relationship is seen for their performance on the Academic Achievement and Reading, Auditory-Perception and Language, and Psychometric Intelligence measures. The DR children's performance is the most consistent on the Academic Achievement and Reading and Psychometric Intelligence variables, and somewhat less so on the Auditory-Perception and Language, Motor, and Underlining Test measures. Both samples of children demonstrate little relationship in their scores over two administrations of the Visual-Perceptual, Tactile-Perceptual, and Right-Left Awareness test measures.

In attempting to account for the observed differences in the consistency and stability between the two groups' performances, one must remember that the subjects are matched on the basis of grade, sex, psychometric intelligence, and retest interval. Therefore, the most

likely sources of variation that could be expected to have affected the children's performances are: effects due to sampling bias, random changes in the children's performance over time, maturational factors, changes in ability due to treatment or experiential factors, poor inter-rater reliability, and possible ceiling effects in some of the data.

The manner in which the two samples of children were selected quite possibly may have contributed to the DR childrens' lesser reliability. Inclusion within the DR sample was contingent upon severely impaired reading ability, while all of the NR children were selected as being average readers. One of the effects of employing a truncated sample (i.e., DR group) is a much greater probability that improvement will occur in the scores of those children relative to the scores of the NR children. Such a "regression-towards-the-mean" phenomena is unintentionally built into the present study as a result of the conservative criteria employed in operationally defining the DR children. In addition, the law of initial values (Wilder, 1950) suggests that the degree of change in a subject's performance is partly a factor of their initial test score, and the relative impact of any intervening treatment effects. Initial test scores that are low in value would be expected to benefit from any intervening treatment effects more so than initial scores that have a

high value. The relatively low stability measures for the subjects in the current study is likely a partial reflection of this effect, although such a relationship was not examined.

An additional likely source of variance results from random effects over time. Random effects are always present in any experimental investigation, and strict adherence to statistical methods of assessing the significance of results remains the best method of controlling such variation.

An additional source of variation thought to affect the childrens' retest performance are the maturational changes, and the skill acquisition that accompany growth. These changes are assumed to occur over the two years intervening between the initial testing and the re-assessment. The nature of the study precludes the examination of such developmental factors and their relative impact upon the childrens' performance on the various ability measures. Even so, it can be assumed that some factors are more susceptible to developmental change than others. Motor abilities represent one such example, and the childrens' performance on such tests would be expected to reflect a large degree of variation on subsequent assessments. Due to the matching controls instituted in this study, particularly on the grade level and retest interval factors, one may assume that any change in development is relatively equated for each group.

The possible influence of treatment or practice effects serves as a likely source of variance that limits the degree of relationship between subjects' scores over assessment periods. Academic remediation programs for the DR children and practice effects on several of the Motor measures (e.g., Name Writing speed (L), Name Writing speed (R)) serve as examples of such an influence. Although it is difficult to delineate precisely the relative effect of such variation on childrens' performance, it can be assumed that each subject was exposed to experiences that may have potentially altered their retest performance. The nature of the data used in this study precluded investigations into each child's life experiences during the retest interval. As a result, more specific conclusions regarding the possible influence of any treatment effects cannot be drawn from this study.

Inter-rater variation always presents a potential source of error whenever studies of the type presented here are attempted. Although more than one examiner was involved in the collection of the data during the different assessment periods, all had received training in the standardized methods of administration to be used for each of the neuropsychological tests. The effects of inter-rater variance upon the two samples' test-retest performance was not examined in the current study; however, the standardized manner in which the test measures were administered would help limit any possible impact due to such variance.

The use of different raters in this study also allows for the possibility that bias on the part of the examiners may have affected subjects' test scores. Although this is a difficult source of variation to partial out of the study the use of many measures that limit the interaction between examiners and children helps protect against this effect. When such interaction was required, every effort was taken to ensure that the examiner administered the tests in a standardized fashion. In addition, all raters were "blind" to the childrens' MAT scores.

Measures that are often interpreted from the perspective of pathognomonic signs rather than levels of performance, most noticeably those measures that assess visual, tactile, and auditory sensation and perceptual ability, can be expected to result in "ceiling" or "floor" effects for the childrens' performance. Given the small range of scores for such test measures, few errors are necessary to dramatically alter the reliability associated with these variables.

Investigation 3

The third investigation of this study is concerned with the degree to which the retest consistency and stability associated with the performance of the DR children resembled that for two heterogeneous clinical samples of children. The results of the Donaghy (1988) study is selected for the comparison of test-retest reliability because of its larger

sample size, resulting from the amalgamation of the subjects from the Brown (1987) and Paniak (1987) studies together. The Brown (1987) subjects serve as the comparison group for the retest stability of the childrens' performances.

In terms of the consistency of the childrens' long-term retest behaviour a much greater relationship is seen between the Donaghy (1988) sample's initial and retest performance than is observed for the DR sample. Even so, direct comparison of the rank ordering of the Pearson-r coefficients suggests a strong degree of agreement in the relative pattern of consistency within abilities between the two groups. In both samples, the largest degree of consistency is observed for the Motor variables, followed by the Academic Achievement and Reading test measures. The smallest consistency is associated with the Visual-Perceptual and Tactile-Perceptual measures. The two samples are dissimilar in ranking the consistency of the Psychometric Intelligence and Auditory-Perception and Language measures. The Donaghy (1988) study found that those subjects demonstrated fair to excellent consistency on the Psychometric Intelligence measures. The DR children perform with poor consistency on these same measures. For the Auditory-Perception and Language measures, while the DR group's performance were found to be among their more consistent test behaviours, the same was not true for the Donaghy (1988) subjects. For these children, their

performance on the Auditory-Perception and Language measures is generally inferior to that demonstrated for the Motor, Psychometric Intelligence, and Academic Achievement and Reading variables.

Similar to the findings of consistency, the degree of stability in the Brown (1987) subjects' performance generally exceeds that for the DR subjects. The vast majority of the variables examined in the Brown (1987) study were rated fair or better in terms of subjects' stability of performance. In contrast, the overwhelming majority of measures for the DR subjects are associated with poor stability. Clearly, the DR subjects' performance over the retest interval is susceptible to confounding factors reducing their stability.

Despite the differences in the degree of stability between the performances of both samples, a moderate relationship is found between the two rank orders of measures' ICC coefficients. For both samples, the most stable performances are observed on the Psychometric Intelligence measures, followed by the Academic Achievement and Reading measures. Similarly, the Tactile-Perceptual variables are associated with the least degree of stability for both samples' retest performances.

Considering only those Visual-Perceptual measures common to both studies, the performance of both groups are generally in agreement. This indicates a degree of

stability for such measures that is less than that observed for the Psychometric Intelligence, Academic Achievement and Reading, and Auditory-Perception and Language ability variables.

In summary, while the degree of stability and consistency in their performance is less than that found for the two clinical samples, the rank order of the DR childrens' resultant Pearson-r and ICC coefficients agree with those found for the Donaghy (1988) and Brown (1987) subjects. Such a finding is not totally unexpected given the large proportion of learning disabled children that comprise the clinical samples.

It is clear, however, that a greater degree of variability is observed in the retest performances of the DR subjects, relative to the Brown (1987) and Donaghy (1988) subjects. In considering why this might be the case one must consider the factors already presented in the discussion of Investigations 1 and 2 (i.e., "regression-towards-the-mean" phenomenon, random variation in performance, maturational factors, inter-rater reliability and bias, treatment effects, and the restricted range of some measurement scores. Each of these factors can be expected to intervene during the retest interval. Since the possible impact that these factors hold for the reliability of the childrens' performances has been discussed for Investigations 1 and 2, the reader can refer to the early

discussion for a complete explanation.

Additional sources of variation that must also be considered are the large variation in retest interval observed for the Brown (1987) and Donaghy (1988) studies, and the specific composition of both clinical samples. The children comprising the clinical samples are heterogeneous in terms of their clinical classifications, although the preponderance are identified as having learning difficulties (Brown, 1987; Donaghy, 1988). Furthermore, these children were all identified subsequent to an initial neuropsychological assessment as being in need for further assessment, reflecting the severity of their difficulties. No attempt is made in this study to compare the qualitative and quantitative aspects of the clinical samples' difficulties with that of the DR children.

The large degree of variability in the test-retest intervals of the clinical samples could have masked any impact that maturational factors may have had on those childrens' retest performances. Brown (1987) reports a mean retest interval of 2.65-years (SD = 1.74-years) for her study, with a range from 1-month to 12-years. Similar retest intervals were reported by Donaghy (1988) (mean = 2.62-years, SD = 1.73-years).

Investigation 4

This last investigation is an attempt to validate the clinical retest reliability of a variety of

neuropsychological measures. The subjects studied consisted of DR and NR children who were initially assessed over two years (Year-0 to Year-2) and again over an additional two year period (Year-2 to Year-4). This discussion focuses first on the relationship between the childrens' consistency of performance over each of the two assessment intervals and second on a similar discussion of the stability of their performances.

While the rank order comparison reveal a similarity between the rank orders resulting from the two retest assessments, some salient differences are in evidence. First of these differences is the large discrepancy between the magnitude of the consistency indices associated with the children's intial and subsequent retest performances. While the majority of the neuropsychological measures are associated with sufficient consistency to warrant ratings of fair or better over the Year-0 to Year-2 comparison, this is certainly not the case for the Year-2 to Year-4 examination. During the latter assessment periods, virtually all of the children's performances are afforded rating of poor consistency. Clearly, the childrens' performances over the second assessment periods were much more susceptible to extraneous influences.

With the above findings in mind, a comparison of the patterns of the childrens' performances within ability area reveals more similarity than dissimilarity. Both samples of

retest behavior are the most consistent on measures of Academic Achievement and Reading, Auditory-Perception and Language, and Psychometric Intelligence (in that order) and least consistent on the Tactile-Perceptual variables. Some variation in how the Right-Left Awareness, and Visual-Perceptual measures are ranked is apparent. Even though the value of the Pearson-r coefficient associated with the measure of Right-Left Awareness is relatively the same over the two intervals of time, this measure occupies a more favourable rank for consistency in the second assessment. Rated 43rd of the 64 measures in the Year-0 to Year-2 assessment, that rank changed to 7th in the Year-2 to Year-4 examination. Similarly, while the measures of Visual-Perceptual ability generally occupy ranks within the lower half of the distribution during the intial retest assessment comparison, they are ranked within the lower end of the upper half of the distribution in the later comparison.

Why such a reduction in the magnitudes of the Pearson-r coefficients occurred over the second retest interval is difficult to account for. All the possible influences mentioned within the discussion of Investigation 1 are equally applicable within the present discussion. Certainly, given the very large proportion (76 percent) of non-significant correlation coefficients found in the Year-2 to Year-4 comparison the effects of random error are particularly suspect.

Examining the retest stability for the childrens' performances over the same time periods, a somewhat different picture of results emerges. The absolute magnitudes of the ICC coefficients found for all of the two comparisons are similar to each other, with virtually all measures in either comparison being rated as poor. However, a much greater number of coefficients in the Year-2 to Year-4 comparison are negative in value. The performance of the children during this time interval suggests that it was worsening over the two years. Indeed, examination of the raw data reveals that this is the case. Why this might be the case is again difficult to explain. Unfortunately, other than the (unspecified) possible effects of factors mentioned previously (i.e., random variation in performance, maturational factors, inter-rater reliability and bias, treatment effects, and possible ceiling effects in some test measurements) little more can be speculated.

The greatest (relative) stability observed for the two comparisons differ in terms of the measures involved. The Year-0 to Year-2 comparison demonstrates the greatest stability for measures of Academic Achievement and Reading, while the Year-2 to Year-4 comparison found that the Tactile-Perceptual variables are the most stable. Both assessment comparisons generally agree with their ranking of the Psychometric Intelligence, Motor, and Right-Left Awareness measures' stability. However, the Year-2 to Year-

4 comparison found the the Underlining Test measures are more stable than the Psychometric Intelligence, Auditory-Perception and Language, Motor, and Academic Achievement and Reading measures. For the Year-0 to Year-2 comparison, the Underlining Test measures are more unstable than any of the other ability measures, save those for Tactile-Perceptual.

To summarize, while the results of Investigation 4 validate the clinical pattern of consistency and stability observed for the pooled sample's performance over time, the large number of non-significant coefficients draws into question the meaningfulness of this relationship. It would appear that for the 42 children involved, the consistency of their performances over two-years is quite variable, depending on which two-year comparison one chooses to examine. The variables within the Visual-Perceptual, Tactile-Perceptual, and Right-Left Awareness abilities are generally found to have poor stability and consistency. Performances on measures of Academic Achievement and Reading are found to have good or better consistency, and only fair stability. The majority of the Auditory-Perception and Language measures demonstrate fair consistency, but all are associated with poor stability. Finally, while some of the variables comprising the Psychometric Intelligence, Underlining Test, and Motor abilities are found to have fair or better consistency, a roughly equal amount are judged to be poor; and, all are rated to have poor stability.

Directions for Future Research

The limitations of the current study suggest some possible future research investigations. First, although it is not examined in the current study, the inter-rater reliability of many of the neuropsychological measures warrants examination, particularly when used with children. While one may assume that such reliability does not differ with the representative sample being examined, such an hypothesis must be tested.

Second, one possible strategy for examining the effect of maturation on at least some of the measures used in the present study (e.g., Motor variables) would be to correlate childrens' standardized scores rather than their raw scores. Standardized scores, such as T-scores, allow for the comparison of a person's test result across age groups by standardizing the raw score relative to the appropriate normative sample's mean and standard deviation. If one can be reasonably confident that children of different diagnostic groups (i.e., a clinical sample and a normative sample of children) mature in a similar progression and at at similar rate, then differences observed should reflect variance other than developmental factors.

Finally, the distinctive clinical sample used in this current study is homogeneous only as far as have a general reading disability was concerned. Recognizing that children can develop seemingly similar difficulties in reading

ability due to very different patterns of neuropsychological dysfunction future studies should attempt to use samples of children that are homogeneous with respect to their neuropsychological patterns of abilities and deficits. While it is evident that few studies of any sort attempt to research the test-retest reliability of neuropsychological measures when used with children, it is clear from the results of the present study that differences in these instruments reliability do exist for different groups of children. It is also evident that utilizing so-called "normal" children to assess the clinical reliability of an instrument is not ideal.

APPENDIX A

TEST-RETEST CORRELATIONS FOR 35 CHRONIC SCHIZOPHRENIC PATIENTS (KLONOFF ET AL., 1970)

TEST-RETEST CORRELATIONS FOR 35 CHRONIC SCHIZOPHRENIC PATIENTS (KLONOFF ET AL., 1970)

Tests	Pearson Correlation Coefficient (r)
Trail Making Test - Part A	.84
Speech Sounds Perception Test	.82
Trail Making Test - Part B	.81
Tactual Performance Test (Time)	.78
Halstead Category Test	.72
Seashore Rhythm Test	.72
Tactual Performance Test (Memory)	.63
Tactual Performance Test (Location)	.49
Finger Tapping Test	.45

Note. Mean test-retest interval equals 52 weeks.

APPENDIX B

TEST-RETEST CORRELATIONS FOR 16 ELDERLY PATIENTS WITH DIFFUSE CEREBROVASCULAR DISEASE, AND 29 CONTROLS (MATARAZZO ET AL., 1974)

TEST-RETEST CORRELATIONS FOR 16 ELDERLY PATIENTS WITH DIFFUSE CEREBROVASCULAR DISEASE, AND 29 CONTROLS (MATARAZZO ET AL., 1974)

Tests	Elderly Patients[a] (Pearson-r)	Male Control[b] (Pearson-r)
Trail Making Test - Part A	.78	.46
Speech Sounds Perception Test	.67	.49
Trail Making Test - Part B	.67	.44
Tactual Performance Test (Time)	.87	.68
Halstead Category Test	.96	.60
Seashore Rhythm Test	.58	.37
Tactual Performance Test (Memory)	.75	.40
Tactual Performance Test (Location)	.72	.53
Finger Tapping Test	.44	.24
Strength of Grip Dominant Hand	.62	.55
Strength of Grip Non-Dominant Hand	.60	.52

[a]Mean test-retest interval equals 12.4 weeks.

[b]Mean test-retest interval equals 20 weeks.

APPENDIX C

TEST-RETEST CORRELATIONS FOR 15 CAROTID ENDARTECTOMY PATIENTS (MATARAZZO ET AL., 1976)

TEST-RETEST CORRELATIONS FOR 15 CAROTID ENDARTECTOMY PATIENTS (MATARAZZO ET AL., 1976)

Tests	Pearson Correlation Coefficient (r)
Trail Making Test - Part A	.77
Speech Sounds Perception Test	.89
Trail Making Test - Part B	.88
Tactual Performance Test (Time)	.93
Halstead Category Test	.82
Seashore Rhythm Test	.74
Tactual Performance Test (Memory)	.34
Tactual Performance Test (Location)	.78
Finger Tapping Test	.83
Strength of Grip Dominant Hand	.65
Strength of Grip Non-dominant Hand	.30

Note. Mean test-retest interval equals 20 weeks.

APPENDIX D

TEST-RETEST CORRELATIONS FOR 91 ALCOHOLIC IN-PATIENTS AND 20 MEDICAL IN-PATIENTS (ECKARDT & MATARAZZO, 1981)

TEST-RETEST CORRELATIONS FOR 91 ALCOHOLIC IN-PATIENTS AND 20 MEDICAL IN-PATIENTS (ECKARDT & MATARAZZO, 1981)

Tests	Alcoholic in-Patients[a] (Pearson-r)	Medical In-Patients[b] (Pearson-r)
Trail Making Test - Part A	.53	.87
Speech Sounds Perception Test	.62	.82
Trail Making Test - Part B	.68	.94
Tactual Performance Test (Time)	.70	.93
Halstead Category Test	.74	.87
Seashore Rhythm Test	.53	.59
Tactual Performance Test (Memory)	.59	.51
Tactual Performance Test (Location)	.53	.57
Finger Tapping Test	.71	.90

[a]Mean test-retest interval equals 16.8 days.

[b]Mean test-retest interval equals 22.9 days.

APPENDIX E

CORRELATIONAL COEFFICIENTS FROM BROWN (1987), PANIAK (1987) and DONAGHY (1988)

CORRELATIONAL COEFFICIENTS FROM BROWN (1987), PANIAK (1987) and DONAGHY (1988)

Test Measures	Brown (1987) ICC	Brown (1987) R_p	Paniak (1987) R_p	Donaghy (1988) R_p
Psychometric Intelligence				
WISC Full-Scale IQ	.82	.82	.73	.81
WISC Verbal IQ	.81	.82	.53	.76
WISC Performance IQ	.75	.75	.79	.77
Vocabulary	.65	.66	.39	.63
Information	.60	.60	.34	.55
Similarities	.52	.52	.27	.46
Comprehension	.46	.46	.38	.45
Arithmetic	.52	.52	.49	.52
Digit Span	.56	.56	.50	.54
Object Assembly	.60	.61	.71	.64
Block Design	.60	.60	.70	.64
Picture Arrangement	.51	.52	.53	.53
Picture Completion	.51	.51	.47	.50
Coding	.48	.48	.59	.51
Academic Achievement				
WRAT Reading	.56	.57	.52	.56
WRAT Spelling	.57	.58	.42	.55
WRAT Arithmetic	.47	.51	.51	.50
Language				
PPVT-IQ	.71	.71	.65	.70
Sentence Memory correct	.71	.71	.58	.67
Auditory Closure correct	.59	.60	.69	.62
Verbal Fluency correct	.54	.54	.45	.51
Speech Sounds correct	.47	.48	.51	.48
Motor (Right-hand Dominant)				
Foot Tapping (R) taps/10"	.49	.51	.71	.55
Foot Tapping (L) taps/10"	.54	.56	.61	.58
Finger Tapping (R) taps/10"	.57	.58	.50	.56
Finger Tapping (L) taps/10"	.60	.60	.40	.55
Grip Strength (R) Kg.	.62	.70	.43	.61
Grip Strength (L) Kg.	.60	.67	.48	.61
Name Writing (R) speed	.18	.18	.17	.17
Name Writing (L) speed	.32	.32	.41	.34

Maze (R) speed	.39	.39	.53	.42
Maze (L) speed	.29	.30	.57	.38
Maze Counter (R) errors	.64	.68	.56	.66
Maze Counter (L) errors	.57	.57	.54	.58
Maze Time (R)	.64	.70	.36	.66
Maze Time (L)	.67	.67	.48	.66
Sensory Tests (Right-hand Dominant)				
Finger Agnosia (R) errors	.31	.33	.59	.41
Finger Agnosia (L) errors	.38	.39	.58	.45
TPT (Dom) time	.29	.29	.55	.34
TPT (NDom) time	.36	.38	.47	.35
TPT (both) time	.40	.44	.45	.44
TPT Location correct	.48	.48	.37	.48
TPT Memory correct	.43	.43	.25	.43
Tactile (R) errors	.32	.34	.28	.32
Tactile (L) errors	.24	.25	.08	.19
Visual (R) errors	.15	.15	-.04	.15
Visual (L) errors	.14	.14	.25	.15
Auditory (R) errors	.29	.31	.00	.27
Auditory (L) errors	.22	.22	-.06	.12
Motor (Left-hand Dominant)				
Foot Tapping (R) taps/10"	.68	.69		.74
Foot Tapping (L) taps/10"	.62	.66		.71
Finger Tapping (R) taps/10"	.62	.64		.58
Finger Tapping (L) taps/10"	.61	.63		.56
Grip Strength (R) Kg.	.52	.58		.62
Grip Strength (L) Kg.	.51	.58		.61
Name Writing (R) speed	.27	.30		.18
Name Writing (L) speed	.21	.22		.23
Maze (R) speed	.02	.02		.26
Maze (L) speed	.18	.19		.37
Maze Counter (R) errors	.62	.63		.62
Maze Counter (L) errors	.56	.56		.57
Maze Time (R)	.71	.72		.72
Maze Time (L)	.26	.27		.29
Sensory Tests (Left-hand Dominant)				
Finger Agnosia (R) errors	.62	.62		.59
Finger Agnosia (L) errors	.69	.69		.69
TPT (Dom) time	.41	.42		.42
TPT (NDom) time	.26	.28		.33
TPT (both) time	.22	.29		.38
TPT Location correct	.21	.22		.30
TPT Memory correct	.26	.26		.36
Tactile (R) errors	.38	.45		.44

Tactile (L) errors	.19	.33		.26
Visual (R) errors	.35	.45		.42
Visual (L) errors	.29	.29		.34
Auditory (R) errors	.42	.46		.60
Auditory (L) errors	.02	.02		.08
Miscellaneous				
Target correct	.69	.69	.64	.68

APPENDIX F

TESTS GROUPED BY ABILITY DOMAIN

TESTS GROUPED BY ABILITY DOMAIN

Psychometric Intelligence

WISC (Wechlser, 1949)

Full-Scale IQ	(FSIQ)
Verbal IQ	(VIQ)
Performance IQ	(PIQ)
Vocabulary subtest	(V)
Information subtest	(I)
Similarities subtest	(S)
Comprehension subtest	(Comp)
Arithmetic subtest	(A)
Digit Span subtest	(DS)
Object Assembly subtest	(OA)
Block Design subtest	(BD)
Picture Arrangement subtest	(PA)
Picture Completion subtest	(PC)
Coding subtest	(Cod)

Academic Achievement and Reading

WRAT (Jastak & Jastak, 1965)

Reading subtest	(WRATR)
Spelling subtest	(WRATS)
Arithmetic subtest	(WRATA)

MAT (Durost, et al., 1971)

Word Knowledge	(MATWk)
Reading	(MATR)
Word Discrimination	(MATWd)

Auditory-Perception and Language

PPVT-IQ Form A (Dunn, 1965)	(PPVTIQ)
Sentence Memory Test (Benton, 1965)	(Senmem)
Auditory Closure Test (Kass, 1964)	(AudClo)
Verbal Fluency Test (Strong)	(VFlu)
Speech-Sounds Perception Test (Reitan & Heineman, 1968)	(Ssper)

Tests for Sensory-Perceptual Disturbances (Reitan, 1966)

Auditory Perception (R and L)	(AudR, AudL)

Visual-Perception

Target Test (Reitan, 1969)	(Target)
Thurstone Reversals Test (Doehring, 1968)	(Reverse)
Children's Word Finding Test: Rhymes (Doehring, 1968)	(Rhymes)
Tests for Sensory-Perceptual Disturbances (Reitan, 1966)	
Visual Perception (R and L)	(VisR, VisL)

Tactile-Perception

Tests for Sensory-Perceptual Disturbances (Reitan, 1966)	
Finger Agnosia (R and L)	(FAgR, FAgL)
Tactile Perception (R and L)	(TacR, TacL)

Motor

Strength of Grip (R and L) (Reitan, 1966)	(GripR, GripL)
Finger Tapping Test (R and L) (Knights & Moule, 1967)	(TapR, TapL)
Foot Tapping Test (R and L) (Knights & Moulde, 1967)	(FTapR, FTapL)
Name Writing Test (R and L) (Reitan, 1966)	(NameR, NameL)
Maze Test (Klove, 1963)	
Maze Time (R and L)	(MazeTR, MazeTL)
Maze Counter (R and L)	(MazeCR, MazeCL)
Maze Speed (R and L)	(MazeSR, MazeSL)

Right-Left Awareness

Right-Left Awareness Test (Piaget, 1928)	
Subtests 1 to 6	(RLAware)

Underlining Test

Underlining Test (Doehring, 1968)	
Subtest 1: Single Number	

Subtest 2: Single Geometric Forms
Subtest 3: Single Nonsense Letter
Subtest 4: Gestalt Figure
Subtest 5: Single Letter
Subtest 6: Single Letter in Syllable Context
Subtest 7: Two Letters
Subtest 8: Sequence of Geometric Form
Subtest 9: Four-Letter Nonsense Syllable, Unpronounceable
Subtest 10: Four-Letter Nonsense Syllable, Pronounceable
Subtest 11: Four-Letter Word
Subtest 12: Unspaced Four-Letter Word
Subtest 13: Single Number

APPENDIX G

DESCRIPTION OF SEVERAL TESTS INCLUDED IN THE NEUROPSYCHOLOGICAL BATTERY

DESCRIPTION OF SEVERAL TESTS INCLUDED IN THE NEUROPSYCHOLOGICAL BATTERY[1]

Tests for Sensory Perceptual Disturbances

Tactile Perception

Subject's ability (without the aid of vision) to perceive unilateral stimulation delivered to the right- and left-sides of the face. Unilateral stimulation is then intersperced with bilateral hand and contralateral hand-face stimulation. Raw score is the total number of errors made for each hand and side of face under all conditions.

Auditory Perception

Subject's ability (without the aid of vision) to perceive unilateral stimulation delivered to the right- and left-ear. Unilateral stimulation is then intersperced with bilateral auditory stimulation. Raw score is the total number of errors made for each ear under all condition.

Visual Perception

Subject's ability to perceive unilateral visual stimulation delivered to the right- and left-eye. Unilateral stimulation is then intersperced with bilateral visual stimulation. Procedure is repeated for the upper, middle, and lower visual fields. Raw score is the total number of errors made for each eye under all conditions.

Finger Agnosia

Subject's ability (without the aid of vision) to determine the finger being touched. All five fingers are stimulated four times in an unsystematic fashion. Each hand is tested in turn. Raw score is the total number of errors made for the fingers on each hand.

Target Test

Visual-spatial patterns are tapped out by the examiner which the subject mechanically reproduced following a period of delay (3 sec.). The patterns increase in complexity as the test progresses. Raw score is the total number of patterns correctly reproduced by the subject.

Speech-sounds Perception Test

Subject is required to underline the correct choice among several written morphemes to match the nonsense syllables he hears. Syllables are presented via a taperecording to the subject. Raw score is the total number of correct responses.

Auditory Closure Test

Subject is required to blend into words 23 progressively longer chains of sound elements presented on tape. Raw score is the total number of correct responses.

Sentence Memory Test

Sentences of progressively increasing length (1 to 26 syllables) are presented to the subject via a taperecording. The subject is required to correctly repeat the spoken sentence. Raw score is the total number of correct responses.

Verbal Fluency

Subject is required to verbally state as many words as he or she can, within 60 seconds, that begins with the letter "p" (as in "pig"). The task is repeated with the letter "c" (as in "cake"). Raw score is the combined total number of words produced.

Name Writing

Subject is required to write his or her complete name with a pencil. Test is administered first to the preferred hand and subsequently to the non-preferred hand. Raw score is the time taken for each hand.

Finger Tapping Test

Subject is required to tap (using only a finger motion) as many times as he or she can within a 10 second period. The preferred hand is tested first (four trials) following which the non-preferred hand is tested. Raw score is the average of the best three out of four trials for each hand.

Foot Tapping Test

Subject is required to tap as many times as he or she can within a 10 second period. The preferred foot is tested first (four trials) following which the non-

preferred foot is tested. Raw score is the average of the best three out of four trials for each foot.

Maze Test

Subject is required to run a stylus through a maze which has the blind alleys filled and is placed at a 70 degree angle. Three scores are obtained: The number of contact with the side of the maze, the total amount of time during which the stylus contacts the side of the maze, and the speed at which the maze is completed (total time from beginning to end). There are two successive trials with the dominant hand, followed by two trials with the non-dominant hand. Raw score is the totals for the two trials with each hand.

Children's Word Finding Test

Rhymes

Subject is presented with five sentences, spoken by the examiner, each of which refer to a specific object. A nonsense word, "Grobnick", replaces the object-to-be-named in each of the sentences. The subject is required to determine the object-to-be-named from the contextual cues presented in the sentences. The test is comprised of 13 such elements. Raw score is the total number of correct responses.

Underlining Test

This test is comprised of 13 subtests. In each subtest the subject is required to visually scan a page of visual elements and underline all of the elements that match a specific stimulus figure. In each subtest, the stimulus figure becomes more complex and requires increased verbal decoding skills. Subtests 1 and 13 present similar types of stimulus figures in order to test for practice effects. Raw score is the net correct for each subtest (correct answers minus incorrect answers).

---

[1] adapted from DeLuca (1986) and Rourke et al. (1986)

REFERENCES

Anastasi, A. (1982). Psychological testing (5th ed.). New York: Macmillan.

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. Psychological Reports, 19, 3-11.

Bartko, J. J. (1976). On various intraclass correlation reliability coefficients. Psychological Bulletin, 83, 762-765.

Bartko, J. J., & Carpenter, W. T. (1976). On the methods and theory of reliability. Journal of Nervous and Mental Disease, 163, 307-317.

Benton, A. L. (1965). Sentence Memory Test. Iowa City, IW: Author.

Berk, R. A. (1979). Generalizability of behaviourial observations: A classification of interobserver agreement and interobserver reliability. American Journal of Mental Deficiency, 83, 460-472.

Bloom, B. S. (1964). Stability and change in human characteristics. New York: John Wiley & Sons.

Bochner, S. (1978). Reliability of the Peabody Picture Vocabulary Test: A review of 32 selected research studies published between 1965 and 1974. Psychology in the Schools, 15, 320-327.

Brown, S. J. (1987). Patterns of reliability of neuropsychological measures in children. Unpublished doctoral dissertation, University of Windsor, Windsor,

Ontario, Canada.

Brown, S. J., Rourke, B. P., & Cicchetti, D. V. (In press). Reliability of tests and measures used in the neuropsychological assessment of children. Clinical Neuropsychologist.

Charter, R. A., Adkins, T. G., Alekoumbides, A., & Seacat, G. F. (1987). Reliability of the WAIS, WMS, and Reitan Battery: Raw scores and standardized scores corrected for age and education. International Journal of Clinical Neuropsychology, 9, 28-32.

Coleman, J. C. (1963). Stability of intelligence test scores in learning disorders. Journal of Clinical Psychology, 19, 295-298.

Conklin, R. C., & Dockrell, W. B. (1967). The predictive validity and stability of WISC scores over a four year period. Psychology in the Schools, 4, 263-266.

Cronbach, L. J., Ikeda, H., & Avner, R. A. (1964). Intraclass correlation as an approximation to the coefficient of generalizability. Psychological Reports, 15, 727-736.

Cronbach, L. J., Rajaratnan, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 16, 137-163.

Dean, R. S. (1980). The use of the Peabody Picture Vocabulary Test with emotionally disturbed adolescents. Journal of School Psychology, 18, 172-175.

DeLuca, J. W. (1986). Identification of subtypes of learning

disabled children with arithmetic disorders: An neuropsychological, multivariate analysis. Unpublished doctoral disseration, University of Windsor, Windsor, Ontario, Canada.

Doehring, D. J. Patterns of impairment in specific reading disability. Bloomington, IN: Indiana University Press.

Dodrill, C. B., & Troupin, A. S. (1975). Effects of repeated administrations of a comprehensive neuropsychological battery among chronic epileptics. Journal of Nervous and Mental Disease, 161, 185-190.

Donaghy, S. (1988). Test-retest reliability of neuropsychological measures in a clinical sample of children. Unpublished manuscript, University of Windsor, Windsor, Ontario, Canada.

Dunn, L. M. (1965). Expanded manual for the Peabody Picture Vocabulary Test. Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Dunn, L. M. (1981). Peabody Picture Vocabulary Test - Revised: Manual for forms L and M. Circle Pines, MN: American Guidance Service.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psychometrika, 16, 407-424.

Eckardt, M. J., & Matarazzo, J. D. (1981). Test-retest reliability of the Halstead Impairment Index in hospitalized alcoholic and nonalcoholic males with mild to moderate neuropsychological impairment. Journal of Clinical

Neuropsychology, 3, 257-269.

Eno, L., & Woehlke, P. (1980). Diagnostic differences between educationally handicapped and learning disabled students. Psychology in the Schools, 17, 469-473.

Estes, B. W. (1955). Influence of socioeconomic status on Wechsler Intelligence Scale for Children: Addendum. Journal of Consulting Psychology, 19, 225-226.

Friedman, R. (1970). The reliability of the Wechsler Intelligence Scale for Children in a group of mentally retarded children. Journal of Clinical Psychology, 26, 338-349.

Gehman, I. H., & Matyas, R. P. (1956). Stability of the WISC and Binet tests. Journal of Consulting Psychology, 20, 150-152.

Green, J. R., Troupin, A. S., Halpern, L. M., Friel, P., & Kanarek, P. (1974). Sulthiame: Evaluation as an anticonvulsant. Epilepsia, 15, 329-349.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley and Sons.

Haggard, E. A. (1958). Intraclass correlation and analysis of variance. New York: Dryden Press.

Holloway, H. D. (1954). Effects of training on the SRA Primary Mental Abilities (Primary) and the WISC. Child Development, 25, 253-263.

Hoyt, C. (1941). Test reliability estimated by analysis of variance. Psychometrika, 6, 153-160.

Irwin, D. O. (1966). Reliability of the Wechsler Intelligence Scale for Children. Journal of Educational Measurement, 3, 287-292.

Jastak, J. F., & Jastak, S. (1978). The Wide Range Achievement Test: Manual of instructions (Revised ed.). Jastak Assessment Systems.

Kass, C. E. (1964). Auditory Closure Test. In J. J. Olson & J. L. Olson (Eds.), Validity studies on the Illinois Test of Psycholinguistic Abilities. Madison, WI: Photo Press.

Knights, R. M., & Moule, A. D. (1967). Normative and reliability data on finger and foot tapping in children. Perceptual and Motor Skills, 25, 717-720.

Klonoff, H., Fibiger, C. H., & Hutton, G. H. (1970). Neuropsychological patterns in chronic schizophrenia. Journal of Nervous and Mental Disease, 150, 291-300.

Lakey, M. A., Downey, R. G., & Saal, F. E. (1983). Intraclass correlations: There's more there than meets the eye. Psychological Bulletin, 93, 586-595.

Lindquist, E. F. (1953). Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin.

Littell, W. M. (1960). The Wechsler Intelligence Scale for Children: Review of a decade of research. Psychological Bulletin, 57, 132-156.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mandel, R., & McLeod, P. (1970). A longitudinal investigation

of the stability of IQ's on the Peabody Picture Vocabulary Test with high and low socioeconomic subjects. Exceptional Children, 37, 300-301.

Matarazzo, J. D., Matarazzo, R. G., Wiens, A. N., Gallo, A. E., & Klonoff, H. (1976). Retest reliability of the Halstead Impairment Index in a normal, a schizophrenic, and two samples of organic patients. Journal of Clinical Psychology, 32, 338-349.

Matarazzo, J. D., Wiens, A. N., Matarazzo, R. G., & Goldstein, S. G. (1974). Psychometric and clinical test-retest reliability of the Halstead Impairment Index in a sample of healthy, young, normal men. Journal of Nervous and Mental Disease, 158, 37-49.

McCue, P. M., Shelly, C., & Goldstein, G. (1986). Intellectual, academic and neuropsychological performance levels in learning disabled adults. Journal of Learning Disabilities, 19, 233-236.

Morrison, M. W., Gregory, R. J., & Paul, J. J. (1979). Reliability of the Finger Tapping Test and a note on sex differences. Perceptual and Motor Skills, 48, 139-142.

Naglieri, J. A., & Parks, J. C. (1980). Wide Range Achievement Test: A one-year stability study. Psychological Reports, 47, 1028-1030.

Nie, N. H. (1983). SPSSx user's guide. New York: McGraw-Hill.

Nunnally, J. C. (1978). Psychometric theory. Toronto: McGraw-Hill.

Paniak, C. E. (1987). Test-retest reliability of neuropsychological measures in a clinical sample of children: A replication attempt. Unpublished manuscript, University of Windsor, Windsor, Ontario, Canada.

Quereshi, M. Y. (1968). Practice effects on the WISC subtest scores and IQ estimates. Journal of Clinical Psychology, 24, 79-85.

Reger, R. (1962). Repeated measurements with the WISC. Psychological Reports, 11, 418.

Reitan, R. M. (1966). Manual for administration of neuropsychological test batteries for adults and children. Seattle, WA: Author.

Reitan, R. M., & Davison, L. A. (1974). Clinical neuropsychology: Current status and applications. Washington, DC: Winston & Sons.

Reitan, R. M., & Heineman, C. (1968). Interactions of neurological deficits and emotional disturbances in children with learning disorders: Methods for their differential assessment. In J. Hellmuth (Ed.), Learning Disorders, Vol. 3., pp. 93-135. Seattle, WA: Special Child Publications.

Reitan, R. M., & Wolfson, D. (1985). The Halstead-Reitan Neuropsychological Test Battery: Theory and clinical interpretation. Tucson, AZ: Neuropsychology Press.

Rourke, B. P. (1966). The effect of anxiety on the development of causal thinking and performance on a cognitive-perceptual task. (Doctoral dissertation, Fordham University), Ann

Arbor, MI: University Microfilms, No. 66-7100.

Rourke, B. P., Fisk, J. L., & Strang, J. D. (1986). Neuropsychological assessment of children: A treatment-orientated approach. New York: Guilford Press.

Rourke, B. P., Ridgley, B. A., & Orr, R. R. (1973). The neuropsychological abilities of normal and retarded readers: A two-year follow-up. Unpublished manuscript, University of Windsor, Windsor, Ontario, Canada.

Russell, D. L., & Rourke, B. P. (1984). Concurrent and predictive validity of level of phonetic accuracy of misspellings for learning disabled children. Unpublished manuscript, University of Windsor, Windsor, Ontario, Canada.

Sarazin, F. F-A., & Spreen, O. (1986). Fifteen-year stability of some neuropsychological tests in learning disabled subjects with and without neurological impairment. Journal of Clinical and Experimental Neuropsychology, 8, 190-200.

Satz, P., Friel, J. (1974). Some predictive antecedents of specific reading disability: A preliminary two-year follow-up. Journal of Learning Disabilities, 7, 437-444.

Sechrest, L. (1984). Reliability and validity. In A. S. Bellack & M. Hersen (Eds.), Research methods in clinical psychology (pp. 24-54). Toronto: Pergamon.

Smith, M. D., & Rogers, C. M. (1978). Reliability of standardized assessment techniques when used with learning disabled children. Learning Disability Quarterly, 1(3), 23-31.

Stevenson, H. W., Parker, T., Wilkinson, A., Hegion, A., & Fish, E. (1976). Longitudinal study of individual differences in cognitive development and scholastic achievement. Journal of Educational Psychology, 68, 377-400.

Strong, R. T. (no date). Verbal Fluency Test. Phoenix, AZ: Author.

Thorndike, R. L. (1940). "Constancy" of the IQ. Psychological Bulletin, 37, 167-186.

Throne, F. M., Schulman, J. L., & Kaspar, J. C. (1962). Reliability and stability of the Wechsler Intelligence Scale for Children for a group of mentally retarded boys. American Journal of Mental Deficiency, 67, 455-457.

Turner, R. K., Mathews, A., & Rachman, S. (1967). The stability of the WISC in a psychiatric group. British Journal of Educational Psychology, 37, 194-200.

Wechsler, D. (1949). Wechsler Intelligence Scale for Children. New York: Psychological Corporation.

Whately, R. G., & Plant, W. T. (1957). The stability of W.I.S.C. IQ's for selected children. Journal of Psychology, 44, 165-167.

Wilson, B. C., Iacoviello, J. M., Wilson, J.J., & Risucci, D. (1982). Purdue Pegboard performance of normal preschool children. Journal of Clinical Neuropsychology, 4, 19-26.

Woodward, C. A., Santa-Barbara, J., & Roberts, R. (1975). Test-retest reliability of the Wide Range Achievement Test. Journal of Clinical Psychology, 31, 81-84.

Zingale, S. A., Smith, M. D., & Dokecki, P. R. (1980). Temporal stability of the Metropolitan Achievement Test when used with learning disabled children. Learning Disability Quarterly, 3(2), 84-86.

VITA AUCTORIS

Michael C. S. Harnadek was born on July 24, 1961 in Winnipeg, Manitoba. In June, 1979 he graduated from Stanley Humphries Secondary School, Castlegar, British Columbia. He enrolled at Selkirk College, Castlegar, British Columbia in September, 1981. He transferred to the University of Victoria in September, 1985; and, graduated in May, 1987 with the Bachelor of Sciences (Honours) degree. Since September 1987 he has been enrolled in the Master's programme in clinical neuropsychology at the University of Windsor.

Michael Harnadek married Gloria M. Grace in August, 1988.